Les contraintes de capacité ferroviaires : une approche économique

Maria Perez Herrero

N° d'ordre NNT : 2016LYSE2162

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 486 Sciences Économiques et de Gestion

Discipline : Sciences économiques

Soutenue publiquement le 12 décembre 2016, par :

María PEREZ HERRERO

Rail Capacity Constraints: an Economic Approach

Devant le jury composé de :

Alain AYONG LE KAMA, Professeur des universités, Université Paris Ouest Nanterre la Défense, Président

Chris NASH, Professeur d'université, University of Leeds, Rapporteur

Erik VERHOEF, Professeur d'université, Université d'Amsterdam, Rapporteur

Grégoire MARLOT, Expert, Examinateur

Julien BRUNEL, Expert, Examinateur

Yves CROZET, Professeur émérite des universités, Université Lumière Lyon 2, Directeur de thèse

UNIVERSITÉ LYON 2

École Doctorale 486 Sciences Économiques et de Gestion

Faculté des Sciences Economiques et de Gestion

Laboratoire Aménagement Economie Transports

Les contraintes de capacité ferroviaires: une approche économique

María Pérez Herrero

Thèse de doctorat de sciences économiques

Soutenance prévue le 12 décembre 2016, devant le jury composé de :

Pr. Chris Nash	University of Leeds	Rapporteur
Pr. Erik Verhoef	Vrije Universiteit Amsterdam	Rapporteur
Pr. Yves Crozet	Université Lyon 2	Directeur de thèse
Dr. Grégoire Marlot	SNCF	Examinateur
Dr. Julien Brunel	SNCF Réseau	Examinateur

Introduction

Les nouvelles formes de mobilité avec une concentration des flux à des heures et des endroits très localisés ont mis en lumière une dégradation de la régularité et donc une augmentation des coûts moyens pour les usagers, à partir d'un certain seuil de circulations.

Les infrastructures de transport sont caractérisées par une capacité fixe à court terme et l'expansion des capacités ferroviaires (infrastructure ou matériel roulant) demandent du temps et de forts investissements. Parallèlement, dans un contexte de restriction budgétaire, les investissements en capacité rencontrent des obstacles financiers depuis plusieurs années. D'un point de vue de l'infrastructure, les investissements dans les nœuds de congestion apparaissent surtout dans et autour des grandes agglomérations, où les coûts de construction sont très élevés (coûts du foncier, nécessité de passer en souterrains, chantiers difficiles d'accès, etc.) : le ferroviaire, un système traditionnellement associé à des rendements d'échelle croissants, semblerait rentrer dans une zone de rendements de densité décroissants pour des augmentations importantes de capacité.

Trouver l'équilibre entre une offre de service ferroviaire et une qualité de service offert, tout en considérant les ressources disponibles, est un des enjeux majeurs pour les gestionnaires d'infrastructure ferroviaire.

La vision des contraintes de capacité dans le monde ferroviaire a été étudiée de façon très compartimentée. D'un côté, il existait une recherche opérationnelle avec comme objectif l'optimisation de la grille horaire d'un point de vue technique. De l'autre, et dans un contexte réglementaire européen, on a constaté un intérêt croissant pour des réflexes économiques, principalement théoriques, en considérant la congestion ferroviaire comme une externalité négative (comme c'est le cas dans d'autres modes de transport) et s'interrogeant sur la pertinence de la tarification comme mesure corrective. En revanche, il n'existait pas de vision globale de la congestion ferroviaire permettant d'articuler les réponses optimales aux désajustements entre l'offre et la demande que génèrent les contraintes de capacité.

Jusqu'à présent, cette question a principalement été étudiée du point de vue de l'ingénierie, dans un univers monopolistique où la répartition de la capacité et les ajustements en cas de conflit étaient gérés par des processus internes. Néanmoins, compte tenu d'une ouverture progressive à la concurrence du monde ferroviaire, analyser économiquement cette question devient un enjeu clé pour le gestionnaire d'infrastructure, dans un contexte de plus en plus régulé.

Cette thèse décrit de façon précise les éléments techniques et les fondements économiques qui permettent de caractériser la problématique de la contrainte de capacité ferroviaire dans son ensemble.

Principales enseignements de cette recherche

Perspectives techniques et économiques

Dans un premier temps, la définition de la contrainte de capacité a été étudiée dans la perspective de l'ingénieur à travers la conception de l'horaire, un élément central de la rencontre entre l'offre et la demande pour les transports programmés. La programmation en amont de l'offre ferroviaire détermine la caractérisation des contraintes de capacité dans le transport ferroviaire.

Un des paramètres clés de la définition de capacité ferroviaire est le niveau de qualité de service souhaité. L'analyse des différents processus de production horaire en France et en Europe montre que les gestionnaires d'infrastructure ferroviaire en sont conscients et intègrent dans leur choix horaires le lien entre la capacité et la robustesse de leur exploitation. Toutes choses égales par ailleurs, il existe un arbitrage entre robustesse du sillon et temps de parcours, ainsi qu'entre robustesse du graphique et capacité.

Néanmoins, on observe une disparité de pratiques entre les gestionnaires, par pays et par type de réseau. De façon générale, les normes de robustesse appliquées pour les différents réseaux sont souvent tacites et fondées sur des retours d'expérience empiriques, les gestionnaires de réseaux semblent procéder par tâtonnement pour déterminer certaines de leurs règles de conception des horaires.

L'analyse de la notion et de la mesure de la contrainte de capacité d'un point de vue économique est un sujet qui a été largement traité dans d'autres modes de transport comme la route ou l'aérien. Une revue de la littérature approfondie des autres modes de transport nous ont permis d'en tirer des enseignements utiles pour la formalisation économique de la problématique ferroviaire (jusqu'à présent, très peu étudiée) et de nuancer leur transposition à l'industrie ferroviaire.

Une fonction de coût généralisé de l'usager

C'est la conjugaison de ces deux visions, la vision technique de l'ingénieur et la vision économique développée dans les autres modes de transport, qui nous a permis d'élaborer un modèle microéconomique du coût généralisé de l'usager, considérant les spécificités ferroviaires de la construction horaire.

Pour cela, le modèle identifie les conséquences pour l'usager de programmer différentes fréquences en termes de services offerts. Du point de vue de l'usager, la contrainte de capacité peut s'exprimer sous deux formes complémentaires, mais non mutuellement exclusives, d'une part le coût de deshorage (effet Mohring) et d'un autre côté l'espérance du coût du retard, lié à un usage intensif du réseau.

Les résultats du modèle permettent de définir le nombre de fréquences optimales qui maximisent le surplus des usagers, en fonction de différents paramètres. Le planificateur recherche une fréquence optimale, sachant que, ceteris paribus, des fréquences élevées diminuent les coûts du deshorage des usagers (les horaires souhaités des usagers seront plus proches des horaires de départ des trains), mais augmentent leur espérance du coût du retard lié à une forte densité de trafic.

La définition de la fonction de coût généralisé de l'usager spécifique au ferroviaire a permis d'objectiver les arbitrages (jusqu'à aujourd'hui tacites) entre la capacité offerte et la qualité de service en termes de fiabilité.

La correction des externalités

Une fois que les coûts pour les usagers ont été déterminés, on a également considéré les coûts opérationnels des entreprises ferroviaires, afin de s'interroger sur les équilibres offre et demande selon les différentes structures de marché, définissant les prix et quantités optimales. Cette analyse a permis ainsi de déterminer sous quelles conditions et avec quels objectifs, les pouvoirs publics (régulateur) doivent intervenir pour ajuster les inefficacités issues des décisions privées des opérateurs. Dans ce contexte, l'analyse développée démontre que, sous certaines conditions, le régulateur peut être amené à valider une tarification de la contrainte de capacité, afin d'internaliser les effets externes générés et envoyer les bons signaux-prix aux agents économiques.

La mise en œuvre d'une tarification de la congestion comme mesure corrective des externalités liées à la contrainte de capacité dépend de plusieurs paramètres :

- **La présence d'un monopole sur le marché** : un monopole rationnel et parfaitement discriminant, détermine lui-même une allocation des capacités de façon efficace et internalise complètement les phénomènes liés à la contrainte de capacité (effet Mohring ou retard). La justification de l'intervention publique dans le cas d'une entreprise monopolistique est liée à l'inefficacité naturelle du monopole (mark-up des prix supérieur au coût marginal) et non à la non-internalisation des contraintes de capacité.

- **La concurrence antre les activités ferroviaires** : une tarification de la congestion ne peut en conséquence qu'être justifiée que dans un contexte de multiples opérateurs (à minima un duopole) sur le marché. Dans la réalité, l'infrastructure ferroviaire supporte différentes activités (transport régional, transport longue distance, transport de fret, etc.) et c'est la combinaison des différentes demandes d'activités qui peut atteindre ou dépasser les capacités du système et entraîner un phénomène de congestion, les unes par rapport aux autres.

  Même si ces différents services sont dans leur majorité fournis par un même opérateur historique, la prise de décision du nombre de circulations demandées par activité se fait de façon non-coordonnée par des instances différentes (autorités organisatrices du transport régional ou national pour les activités conventionnes, et par les opérateurs pour les activités commerciales). Dans le cas de l'infrastructure ferroviaire, la concurrence entre activités peut également être considérée à l'origine d'externalités.

- **Le pouvoir de marche des opérateurs** : comme dans le secteur aérien, une tarification de la congestion justifiée par la non considération des externalités, doit être minorée en fonction du pouvoir de marché des entreprises. En effet, un certain pouvoir de marché, lié par exemple au faible niveau de concurrence intermodale, donne aux opérateurs l'opportunité de discriminer via ses prix. Si la tarification de la congestion faisait abstraction de ce pouvoir de marché, sa mise en œuvre créerait une distorsion supplémentaire dans le marché.

- **Le niveau de qualité initial, en termes de fréquences offertes et fiabilité** : une fréquence supplémentaire sur une ligne n'engendre pas la même externalité en fonction de la typologie des incidents à l'origine (niveau retard initial) ou de la fréquence initialement offerte. A même niveau de fréquences, une ligne avec une probabilité d'incident plus élevée générerait une externalité supérieure à celle

d'une ligne avec un niveau de fiabilité important. Dans certains cas, une fréquence additionnelle génère une externalité positive (effet Mohring), justifiant ainsi une subvention pour inciter la production supplémentaire et non une tarification complémentaire.

Une perspective plus large de la gestion optimale de la contrainte de capacité

Dans cette approche, le bon signal-prix qui découle de cette analyse complète se base sur une vision statique de la question, et n'est pertinent que dans une vision instantanée, à court terme. Il est pourtant nécessaire de souligner que les pouvoirs publics (ou le régulateur) ne doivent pas isoler la question de la tarification de la congestion des autres composantes du problème des contraintes de capacité. La tarification proposée dans cette recherche, qui constituerait un outil optimal pour résoudre les inefficacités, se fonde sur le postulat selon lequel le dimensionnement et le niveau de fiabilité du réseau sont optimaux et fixes à court terme. L'analyse globale des contraintes de capacité doit s'inscrire dans une vision de long terme, incluant le coût de développement de la capacité, et également considérer la variation de ses paramètres et de son impact sur les recommandations tarifaires.

Il ressort de ce travail que le régulateur (au sens large) ne doit pas soutenir une tarification de la congestion sans s'assurer que le gestionnaire d'infrastructure alloue la capacité de la façon la plus efficace possible. Une tarification de la congestion telle que décrite précédemment est assujettie à une allocation optimale des capacités, fondée sur une connaissance fine de la fonction de coût généralisé de l'usager par les opérateurs/gestionnaire d'infrastructure ainsi que de leurs propres coûts. Si ces conditions ne sont pas respectées (mauvaise allocation des capacités par méconnaissance de la fonction des coûts des usagers), la mise en œuvre de la tarification de la congestion peut conduire à une situation inoptimale, et entraîner une perte de valeur pour la collectivité. Imaginons par exemple un gestionnaire d'infrastructure très averse aux retards, qui néglige la valeur pour l'usager de la fréquence, et qui surestime les marges horaires, en restreignant la capacité. Si le régulateur décide d'autoriser une tarification de la congestion, une augmentation des prix inciterait à une réduction de fréquences demandées en deçà de la fréquence optimale. Une tarification de la congestion dans un contexte de surestimation des marges se traduirait par une sous-utilisation de la capacité optimale disponible.

En somme, la présente recherche met l'accent sur l'importance des composantes de

la fonction de coût généralisé de l'usager. Avoir une connaissance fine de cette fonction est nécessaire pour que le gestionnaire d'infrastructure objective les arbitrages réalisés dans le cadre du processus de construction horaire, mais aussi de justifier la pertinence et l'optimalité d'une éventuelle tarification de la congestion à la puissance publique.

Néanmoins, pour que les préconisations à court terme évoquées restent optimales dans une perspective économique de long-terme, l'inadéquation entre l'offre et la demande liée à la congestion doit être considérée à l'échelle du système ferroviaire, c'est-à-dire en comparant, d'un côté la demande finale (voyageurs ou marchandises) et, de l'autre, l'offre ferroviaire en termes de places offertes, de fréquences, de fiabilité et de dimensionnement.

Vers une mise en œuvre de la tarification dans l'actuel cadre réglementaire

L'intuition économique plaidant pour une vision d'ensemble dans l'analyse des contraintes de capacité se retrouve dans le cadre juridique européen et national du système ferroviaire. La directive européenne 2012/34 détermine le cadre légal du lien entre ces paramètres de façon explicite.

Le point de vue européen

Comme décrit dans cette thèse, sous certaines conditions, la mise en œuvre d'une tarification de la congestion peut être considérée comme une mesure corrective des externalités liées à la contrainte de capacité. Ainsi, l'article 31 de la directive européenne autorise le gestionnaire d'infrastructure à appliquer *"une redevance au titre de la rareté des capacités de la section identifiable de l'infrastructure pendant les périodes de saturation"*. Dans ce cadre réglementaire, une tarification liée à la contrainte de capacité peut intervenir si le gestionnaire d'infrastructure a, au préalable, formellement déclaré saturée une ligne ou section de ligne de l'infrastructure.

L'article 47 de cette directive décrit les dispositions réglementaires concernant la saturation de l'infrastructure. Pour tenir compte de la transposition de la directive 2012-34, l'article 26 du décret no.2003-194 modifié en août 2015 considère que le gestionnaire de l'infrastructure doit déclarer une section de l'infrastructure comme saturée *"lorsque le gestionnaire d'infrastructure constate, à l'issue de la procédure de programmation et de coordination des capacités et de la consultation des candidats, l'impossibilité de répondre*

favorablement à toutes les demandes de capacité sur une section de l'infrastructure pendant certaines périodes (...)". La réglementation ajoute que *"Il en va de même pour des sections susceptibles de souffrir d'une même pénurie dans un proche avenir"*. La particularité de ce processus tel que transposé en droit national par rapport au droit européen est que la saturation peut être *"constatée"*ou *"prévisible"*.

Selon l'article 50 de la directive européenne, la déclaration de saturation doit être suivie d'une analyse de la capacité dans un délai de 6 mois. Cette analyse a pour objectif de déterminer les raisons de la saturation et proposer des mesures correctives pour y remédier. Selon l'interprétation tout au long de cette recherche, cette analyse de capacité devrait permettre au gestionnaire d'infrastructure de justifier à ce moment ses arbitrages entre fréquence et régularité et prouver par exemple que si un sillon est refusé, c'est dans l'intérêt de la collectivité, afin de ne pas dégrader un certain niveau de qualité de service. D'un point de vue économique, la justification d'une réponse non-favorable aux demandes de capacité devrait être établie sur la possibilité de démontrer qu'il existe une allocation optimale préalable des capacités disponibles.

Enfin, comme l'énonce l'article 51 de la directive, dans les six mois suivants l'analyse des capacités, le gestionnaire d'infrastructure doit proposer un plan de renforcement de ces dernières qui peut être soumis à l'approbation de l'État. Ce plan doit définir *"les raisons de la saturation, l'évolution probable du trafic, les contraintes qui pèsent sur le développement de l'infrastructure ainsi que les solutions envisageables concernant le renforcement des capacités"*. L'application d'une redevance supplémentaire pendant les périodes de saturation est assujettie à la présentation et mise en œuvre des actions définies dans le plan de renforcement.

D'un point de vue réglementaire, le gestionnaire d'infrastructure doit renoncer à percevoir la redevance s'il ne met pas en œuvre les actions du plan de renforcement cité. Ainsi, la directive établit un lien entre les recettes supplémentaires perçues par le gestionnaire d'infrastructure pendant les périodes de saturation et une politique de renforcement de la capacité (dans laquelle, on pourrait trouver des investissements en capacité, mais pas uniquement). Elle établit ainsi une relation entre les instruments de régulation à court (tarification), et long terme (investissement, et autres mesures de renforcement) des contraintes de capacité ferroviaires.

Le cas français

Même si le cadre légal offre la possibilité de déclarer des infrastructures saturées, le gestionnaire d'infrastructure français n'a pour l'heure jamais mis en œuvre cette possibilité. Jusqu'à présent, la mise en place d'une procédure de déclaration de saturation avait été considérée complexe d'un point de vue pratique et superflue car le processus de coordination permettait de résoudre tous les conflits. Toutefois, certains axes et nœuds (i.e. la ligne à grande vitesse entre Paris et Lyon et le nœud ferroviaire lyonnais (NFL), l'axe Montpellier-Perpignan ou l'axe Marseille-Nice) sont susceptibles d'être saturés à moyen/long terme, si des actions ne sont pas mises en œuvre.

Dans le cadre de régulation français, l'ARAFER, dans son avis no. 2016-014 relatif au DRR pour l'HDS 2017 [1], a recommandé à SNCF Réseau d'utiliser la procédure de déclaration de saturation prévue dans la directive 2012/34/UE et sa transposition en droit national lorsque cela est pertinent.

SNCF Réseau travaille depuis sur la possibilité de mettre en œuvre la procédure de déclaration de saturation prévue par la directive et la tarification qui pourrait être associée, pour l'année de service 2018. La proposition soumise au régulateur par SNCF Réseau inscrit la procédure de déclaration de saturation dans le calendrier du processus d'allocation de capacité en vigueur et fait la distinction de façon claire entre une déclaration de saturation prévisible et constatée. Elle propose également, à terme, un dispositif tarifaire forfaitaire en cas de déclaration de saturation prévisible, visant à inciter les demandeurs de sillons à des changements de comportements. Ce dispositif (dans sa dimension tarifaire) sera proposé à blanc pour la première année (HDS 2018).

Plus généralement, dans les autres pays européens, on observe que la moitié des GI ferroviaires réalisent des déclarations de saturation. En revanche, le périmètre de sections impacté par la déclaration de saturation diffère entre pays et, jusqu'à présent, la mise en place d'une redevance supplémentaire associée à la contrainte de capacité a été peu mobilisée comme levier d'action.

Comme constaté, les réflexions actuelles concernant la mise en œuvre d'une procédure de déclaration de saturation et sa tarification associée se trouvent aujourd'hui dans une étape préliminaire au sein des gestionnaires d'infrastructures européens. Dans le cas français, ces réflexions participent à une démarche progressive de volonté de trans-

[1] Dans ses avis précédents no. 2012-005 et no. 2013-002 relatifs aux DRR 2013 et 2014 et no. 2014-001 et no. 2015-003 relatif aux DRR 2015 et 2016, l'Autorité avait déjà recommandé à SNCF Réseau d'utiliser cette procédure.

parence et objectivation d'un point de vue économique des activités liées aux contraintes de capacité de la part de SNCF Réseau.

Le cadre légal européen inscrit dans une logique d'ouverture à la concurrence et les demandes du régulateur de clarification et justification des procédures inviteront sans doute SNCF Réseau à approfondir ces réflexions et considérer l'ensemble des enjeux et incitations associés dans les années à venir. Dans ce contexte, il semble important d'alimenter le débat avec quelques premiers éléments d'analyse supplémentaire. Ces éléments laissent présager que l'évaluation du coût de la contrainte de capacité, sous le regard d'un régulateur, pourrait traverser les différents domaines de l'accès au réseau (conception de l'infrastructure, allocation des sillons, tarification).

Recommandations

Tout d'abord, les réflexions sur les politiques tarifaires devraient être associées à une réflexion de long terme, incluant une réflexion sur le bon dimensionnement de l'infrastructure. De fait, la validation d'une tarification de la congestion sans logique de politique d'investissements futurs pourrait par exemple inciter le gestionnaire à un sous-investissement en capacité physique. On peut s'interroger si, dans son propre intérêt, le gestionnaire d'infrastructure investirait en capacité si cela signifie une dépense supplémentaire et une diminution de ses recettes de congestion. Il semble indispensable que la régulation aborde la problématique de la contrainte de capacité en considérant l'ensemble des composants et des horizons temps, afin d'assurer un système vertueux d'incitations. Par exemple, pour les entreprises ferroviaires, l'information sur l'existence de périodes de saturation doit intervenir suffisamment tôt pour que l'incitation soit effective, pour qu'elle puisse éventuellement ajuster leur demande de capacité à la présence d'un péage de congestion.

L'optimum de court terme

La théorie économique a abondamment étudié le sujet du lien entre une tarification optimale à court terme et les investissements de capacité à long-terme. Comme décrit dans le Chapitre 2 de ce manuscrit, sous certaines conditions, à l'optimum, les recettes de congestion permettent de couvrir la totalité des dépenses associées à des investissements en capacité. Mise à part la question de l'autofinancement des investissements en capacité (assujetti à des conditions restrictives et peu adaptées à la réalité ferroviaire), il est in-

téressant de regarder en détail le lien entre ces deux variables sous un prisme économique.

L'optimisation de l'arbitrage entre fiabilité et offre ferroviaire conduit à un optimum de court terme : le coût global de la congestion est minimisé et le surplus des usagers est maximisé. Pour autant, cet optimum de court terme comporte une part de congestion résiduelle (usagers ne voyageant pas à l'heure souhaitée ou des voyageurs retardés par la congestion de l'infrastructure). Cette congestion est optimale à court terme, mais la réalisation d'un optimum de long terme suppose que le coût de cette congestion résiduelle demeure inférieur au coût d'une augmentation de capacité de l'infrastructure qui permettrait de la réduire.

L'augmentation de capacité

En effet, une tarification de la congestion engendre des recettes à court terme qui serviront à financer (tout ou partie des investissements en capacité future). A l'optimum, il conviendrait de réaliser un investissement d'accroissement de la capacité quand le coût marginal de la congestion (à capacité donnée) est supérieur au coût marginal d'augmentation de la capacité. Cette condition pose deux questions :

En premier lieu, elle suppose que le gestionnaire de réseau (ou le régulateur) est capable de valoriser le coût de la capacité dans les évaluations socio-économiques préalables. Or on sait que, jusqu'à présent, l'évaluation socio-économique réalisée dans le secteur des transports reposait de façon principale sur la valeur des gains de temps. Ce cadre méthodologique a été particulièrement bien adapté pour justifier de la réalisation des autoroutes, puis des lignes ferroviaire. Il s'est en revanche heurté à certaines difficultés pour valoriser les investissements de création de capacité, même si, récemment, un changement méthodologique (Rapport Quinet, 2013) semble s'amorcer en offrant de nouveaux outils pour valoriser ces projets de développement.

En second lieu, cette condition conduit nécessairement à des déséquilibres temporaires car le monde réel est caractérisé par des indivisibilités. Comme le signale Hau (1998), la séquence optimale du processus décisionnel est d'abord d'établir une politique de mise en œuvre d'une tarification au coût marginal social et ensuite de planifier les ajustements futurs en capacité par rapport à la demande future et les politiques de prix établies. Dans le cas d'un système avec indivisibilités comme le ferroviaire, les ajustements entre la tarification et l'investissement ne se feront pas de façon automatique. Dans un premier temps, il est possible que, compte tenu de la différence du temps entre la prise de décision d'un investissement et sa réalisation effective, la tarification de la congestion interviendra

malgré l'existence d'investissements de capacité aux coûts inférieurs à celui de la capacité résiduelle. De façon symétrique, après la réalisation de l'investissement, le nouveau dimensionnement de la capacité pourrait éliminer la congestion résiduelle qui justifiait une tarification de la congestion, et elle devrait être supprimée. Pour envoyer des signaux prix cohérents sur longue période aux acteurs, la présence de déséquilibres de ce type peut plaider pour une forme de lissage de la tarification de la congestion : le péage serait alors réduit pendant la période de sous-dimensionnement de l'infrastructure, à condition de pouvoir être prolongé une fois l'investissement réalisé.

Néanmoins, un lissage de la tarification de la congestion peut interférer avec l'objectif du bon signal-prix à court terme si les opérateurs ne peuvent plus clairement identifier le coût des externalités générées par leurs décisions privées. Trouver un équilibre entre l'effet incitatif de la tarification pour un usage optimal de l'infrastructure à court terme et sa faisabilité technique dans un monde ferroviaire avec une durée de vie des investissements élevée est une question ouverte et complexe qui demande une réflexion approfondie à part entière.

D'autres composants liés aux contrainte de capacité

De plus, d'autres composantes comme la performance et les efforts liés à son amélioration qui avaient été considérés fixes jusqu'à maintenant doivent être intégrés dans une logique générale de long terme. Comme pour les investissements, cette logique est aussi déterminée par les textes réglementaires. Les considérants de la directive 2012/34/UE disposent que *"il est souhaitable que les entreprises ferroviaires et le gestionnaire de l'infrastructure soient encouragés à réduire au minimum les défaillances et à améliorer les performances du réseau ferroviaire"*. Ainsi, l'article 35 de la directive précise qu'un système d'amélioration de performance *"peut comporter des sanctions en cas d'actes à l'origine des défaillances du réseau, des compensations pour les entreprises qui sont victimes de ces défaillances et des primes en cas de bonnes performances dépassant les prévisions"*.

Dans le cadre de notre analyse, la capacité est considérée allouée de façon efficace à court terme compte tenu des incidents initiaux observés. Néanmoins, si le gestionnaire d'infrastructure et les entreprises ferroviaires ne font pas les efforts nécessaires pour minimiser le nombre d'incidents dans le long terme, le résultat des politiques à courte échéance ne sera pas efficace. Pour cette raison, il faut faire le lien avec le système d'améliorations de la performance, qui existe pour inciter à une réduction des incidents, qu'ils concernent l'infrastructure ou le matériel roulant. Les améliorations de performance peuvent

se traduire par un bon niveau d'investissements dans la fiabilité du réseau ou par une meilleure efficacité du gestionnaire d'infrastructure et des entreprises ferroviaires dans le traitement des incidents.

Vers une vision systémique de la contrainte de capacité ferroviaire

Pour conclure, la congestion ferroviaire ne doit pas être réduite à une relation entre le nombre de trains et les retards ou la mise en place d'une tarification. Dans un monde ferroviaire de plus en plus régulé, avec des processus de décision plus ouverts et concertés, l'analyse des contraintes de capacité doit être le résultat d'une analyse du système, liant toutes les décisions relatives aux capacités à court et long terme comme:

- Grands arbitrages du processus d'allocation de capacité entre fréquence et régularité.
- Les enjeux d'une tarification de la congestion
- Les mécanismes d'incitation à l'amélioration de la régularité (Système d'amélioration de la performance).
- La définition du bon niveau d'investissement en capacité
- La valorisation de projets qui créent de la capacité et/ou permettent d'améliorer la robustesse du graphique dans l'analyse socio-économique.

UNIVERSITÉ
LUMIÈRE
LYON 2

N° d'ordre NNT : 2016LYSE2162

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 486 Sciences Économiques et de Gestion

Discipline : Sciences économiques

Soutenue publiquement le 12 décembre 2016, par :

María PEREZ HERRERO

Rail Capacity Constraints: an Economic Approach

Devant le jury composé de :

Alain AYONG LE KAMA, Professeur des universités, Université Paris Ouest Nanterre la Défense, Président

Chris NASH, Professeur d'université, University of Leeds, Rapporteur

Erik VERHOEF, Professeur d'université, Université d'Amsterdam, Rapporteur

Grégoire MARLOT, Expert, Examinateur

Julien BRUNEL, Expert, Examinateur

Yves CROZET, Professeur émérite des universités, Université Lumière Lyon 2, Directeur de thèse

UNIVERSITÉ LYON 2

École Doctorale 486 Sciences Économiques et de Gestion

Faculté des Sciences Economiques et de Gestion

Laboratoire Aménagement Economie Transports

Rail Capacity Constraints: an Economic Approach

María Pérez Herrero

Thèse de doctorat de sciences économiques

Soutenance prévue le 12 décembre 2016, devant le jury composé de :

Pr. Chris Nash	University of Leeds	Rapporteur
Pr. Erik Verhoef	Vrije Universiteit Amsterdam	Rapporteur
Pr. Yves Crozet	Université Lyon 2	Directeur de thèse
Dr. Grégoire Marlot	SNCF	Examinateur
Dr. Julien Brunel	SNCF Réseau	Examinateur

Caminante, son tus huellas

el camino, y nada más;

caminante, no hay camino,

se hace camino al andar [1]

Antonio Machado

[1] Traveller, the road is only
your footprint, and no more;
traveller, there's no road,
the road is your travelling (Antonio Machado).

Acknowledgements

This dissertation would not have been possible without the help and guidance of several people to whom I would like to express my gratitude.

First and foremost, I am deeply grateful to Prof. Yves Crozet, my PhD supervisor. I would like to thank him for his availability, support and help. I had the opportunity during the last three years to benefit from his insight, encouragement and ideas.

I also thank Prof. Chris Nash and Prof. Erik Verhoef who kindly agreeing to be the rapporteurs for this dissertation.

I want to express my gratitude to Grégoire Marlot and Julien Brunel for pushing me to go a step further in my professional career doing this PhD research. I sincerely thank them for their availability in giving me much guidance along the whole course of this thesis. I also thank Jean-François Ducoing, who allowed me to finish my thesis in good conditions at SNCF Réseau, which funded my research.

I feel very lucky to have had the opportunity to realise this thesis in the Regulation Department of SNCF Réseau. I would like to thank all my colleagues for providing me a stimulating research environment. Special thanks to Amaury, Benjamin, Camille, Carlos, Catherine, Coraline, Corine, Cécile, Damien, Délia, Elise, Jean, Juliette, Laure, Lucie, Mickaël, Sandrine, Sophie, Tamires, Thomas, Philippe, Reinhard, Romina et Xavier for all the coffees we shared, their kind support throughout these years and the fruitful discussions we had together.

I thanks Vicente Montesinos, Prof. of the Universidad Politécnica de Valencia (Spain) for his helpful and accurate mathematical remarks at the take-off of this research.

I thanks Nicolas Coulombel for the attention he brought to my work and for his valuable comments which helped me improve the final version of this dissertation. The third chapter will hopefully be reflected in a peer-review co-written paper.

I would also like to thank all the people of the LAET, specially Morgane Deplanque and Sophia El Bahi for their availability and help in the distance during these years.

I extend my gratitude to all my friends from Valencia, Paris and Strasbourg for their encouragements and friendly support during this (sometimes long) years.

Me gustaría dar las gracias a toda mi familia, de Valencia y Strasbourg, que a pesar de la distancia, siempre han estado cerca. Especialmente a mis padres, Pablo, Marta y mis sobrinos, por su cariño y su apoyo en cada una de mis decisiones. Gracias papás por haber confiado siempre en el valor del capital humano (y en mí).

Enfin, "gracias" Manu, pour ton support sans faille. Merci d'avoir été une source d'énergie et de motivation depuis le début et jusqu'à la fin de cette aventure. Nous tournons une page : vivement la suite!

Abstract

This PhD dissertation addresses the foundations of a detailed characterisation of rail capacity constraints from an economic perspective.

Traditionally, railway capacity has been studied from the standpoint of engineering in a monopolistic world where capacity choices were considered as an organisational issue and set out in internal procedures. However, there is now a growing interest in analysing this issue from an economic perspective, specially regarding the ongoing deregulation tendency.

Firstly, the definition of railway capacity constraints is presented from an engineering perspective via timetable design, a key element in matching supply and demand for planned transport services. A better understanding of timetable construction methods led to highlighting the implicit trade-offs between the capacity supplied and service quality in terms of reliability in the current graphic timetable construction processes in European infrastructure managers. Secondly, this technical vision of the engineer is combined with the economic vision developed for other modes of transport. It allows us to formulate a microeconomic model of the consumer generalised cost function, specific to the railway services. This model highlights the dual effects for the users of a higher frequency of rail traffic. It impacts the expected scheduled delay cost (Mohring effect) on the one hand, and a congestion effect linked to the intensive use of the network on the other. Once the detailed generalised cost function for train users has been determined, we develop an equilibrium model, by considering users' behavior, operators' costs and by describing how supply and demand interact under different market conditions. We analyse the interactions between demand and supply and show that, under some conditions, it is optimal from a welfare point of view to charge the cost of capacity constraints in order to internalize the negative external effects generated, and send the right price signals to economic operators. Nevertheless, in certain cases, an additional frequency generates a positive externality (Mohring effect), thereby justifying a subsidy to encourage using the railway line rather than increases access charges.

Keywords: Capacity constraint; Congestion; Externality; Mohring effect; Railway transport; Regulation; Reliability; Timetable.

Résumé

Cette thèse décrit de façon précise les éléments techniques et les fondements économiques qui permettent de caractériser la problématique de la contrainte de capacité ferroviaire dans son ensemble.

Jusqu'à présent, la question de la contrainte de capacité ferroviaire a principalement été étudiée d'un point de vue ingénierie, dans un univers monopolistique où la répartition de la capacité et les ajustements en cas de conflit étaient gérés par des processus internes. Néanmoins, compte tenu d'une ouverture progressive à la concurrence du monde ferroviaire, analyser économiquement cette question devient un enjeu clé pour le gestionnaire d'infrastructure, dans un contexte de plus en plus régulé.

Ce manuscrit aborde dans un premier temps, la définition de la contrainte de capacité selon la perspective de l'ingénieur, à travers la conception de l'horaire, un élément majeur de la rencontre entre l'offre et la demande pour les transports programmés. Une meilleure connaissance des méthodes de construction horaire a permis de mettre en évidence les arbitrages implicites entre la capacité offerte et la qualité de service en termes de fiabilité. La vision technique de l'ingénieur combinée à la vision économique développée dans les autres modes de transport, nous a permis d'élaborer dans un second temps, un modèle microéconomique du coût généralisé de l'usager, considérant les spécificités ferroviaires de la construction horaire. Cette modélisation a mis en évidence le double effet d'une fréquence ferroviaire supplémentaire, d'une part sur le coût de « deshorage » (effet Mohring) et d'autre part sur l'espérance du coût du retard, lié à un usage intensif du réseau. Une fois la fonction de coût généralisé spécifique au ferroviaire déterminée, nous avons construit un modèle d'équilibre offre-demande, en considérant le comportement des usagers ainsi que les coûts des opérateurs. Ce modèle décrit les interactions entre l'offre et la demande selon les différentes structures de marché. L'analyse développée démontre que sous certaines conditions, le régulateur peut être amené à valider une tarification de la contrainte de capacité, afin d'internaliser les effets externes générés et d'envoyer les bons signaux-prix aux agents économiques. Néanmoins, dans certains cas, une fréquence additionnelle génère une externalité positive (effet Mohring), justifiant ainsi une subvention pour intensifier l'usage de la ligne et non une tarification complémentaire.

Mots clés : Construction horaire ; Contrainte de capacité ; Congestion ; Effet Mohring ; Externalité ; Transport ferroviaire ; Régulation ; Régularité.

Contents

Introduction

Introduction

The study of railway capacity constraints responds to a contemporary challenge for railway infrastructure managers and every stakeholder in the railway industry. Indeed, the change in mobility behaviours and the concentration of activities and population around large cities has led to the polarisation of railway use in recent decades. These changes have increased the number of trips and intensified the use of infrastructures in certain localised parts of the network and during specific hours.

Transport demand requires transport networks which are characterised by a certain fixed capacity in the short term. The increase of capacity (infrastructure and rolling stock) in the railway industry demands time. The investment process for infrastructure managers (IM) and train operating companies (TOC) is long term and covers many years. If the level of demand approaches the limit of infrastructure capacity in the short term, a reduction of service quality may occur.

Service quality, in terms of regularity, has also become a central issue for the attractiveness and the efficiency of railways in dense areas. If the service quality (regularity) of a given infrastructure is considered as fixed, capacity constraint will be expressed exclusively in the form of absolute scarcity. On the contrary, if the goal is to satisfy all the demand, then capacity constraint will be expressed only by the deterioration of regularity. Lastly, if speed or regularity can be degraded without it being vital to serve the entire demand, the capacity constraint will be expressed by these two forms simultaneously.

Increasing the capacity of the network to manage the limits of rail capacity is an option that requires evaluation in the long term. Whatever the case, expanding railway infrastructure capacities as a natural answer to capacity constraints is a long and expensive solution; the sunk costs are high and have come up against public budgetary constraints in the last decade.

Railway capacity is therefore a scarce resource in both time and space. In a framework in which massive investments cannot be contemplated as a realistic solution due to public

budgetary and environmental restrictions, it appears particularly relevant to consider and optimize alternative short-term measures. The debate on whether to take short and long term measures has been discussed extensively and at great length in the academic literature on road congestion. The contemporary railway context has highlighted the importance of understanding and studying rail capacity constraints from the economic and regulatory perspective.

Finding a balance between the demand for rail services and service quality supplied, in terms of regularity, in a world of limited financial resources has now become a major challenge for railway infrastructure managers.

This PhD dissertation focuses on the detailed characterisation of rail capacity constraints seen from an economic perspective. In-depth and global understanding of the issue of rail capacity constraints is necessary to correctly design strategic solutions. Traditionally, railway capacity has been studied from the standpoint of engineering in a monopolistic world where capacity choices were considered purely as an organisational issue and set out in internal procedures. However, there is now growing interest in analysing this issue from the economic perspective in an increasingly regulated environment, in the framework of stiffer competition in the market for infrastructure capacity. The purpose of this research is to reduce this gap between the disciplines of engineering and economics, by proposing tools and insights to analyse capacity constraints in the railway sector.

Improving knowledge on railway capacity constraints would allow IMs to justify to candidate operators and regulatory authorities that capacity is allocated on the basis of transparent and non-discriminatory conditions. It would also allow IMs to define and implement the short-term measures best adapted for optimizing infrastructure utilisation. Furthermore, seen in the long term, it could contribute to orientating capacity investments to include cost-benefit analyses.

A Few Facts on Transport Capacity Constraints

In the last few years, there has been serious and growing concern about the degradation of service quality in all transports services and how it is measured. Well-built databases are indispensable for analysing the evolution of service quality and carry out exhaustive comparative benchmarking between cities around the world.

For example, many databases deal with traffic jams and delays in the road sector. The INRIX National Traffic Scorecard Annual Report (Inrix, 2015) has analysed and compared the status of traffic delays in countries and major metropolitan areas worldwide since 2007.

Based on the average annual hours wasted in traffic, the top 10 most congested cities in Europe in 2015 were:

Europe city rank 2015	Metropolitan area	Hours wasted in traffic 2015
1	London commute zone, UK	101
2	Stuttgart, Germany	73
3	Antwerp, Belgium	71
4	Cologne, Germany	71
5	Brussels, Belgium	70
6	Moscow, Russia	57
7	Karlsruhe, Germany	54
8	Munich, Germany	53
9	Utrecht, Netherlands	53
10	Milan, Italy	52

Table 1: The top 10 most congested cities in Europe in 2014. Source: Inrix (2015)

EUROCONTROL (Eurocontrol, 2016) records and analyses annually delay and cancellations for all-causes for air transport in Europe. Available statistics show that service quality in airports (measured by delays) also represents a considerable time cost for users and an additional operational cost for airline companies. Based on the average delay per flight, the top 10 arrival airports affected by delays are given in Table 2.

In the railway sector, the goal of building a common European database on delays is much more recent. Nevertheless, since 2007, the European Commission has collected data on rail market developments in Member States via RMMS Questionnaires (Commission, 2014). More recently, the European platform of network infrastructure managers (PRIME) has also worked on the construction of a common database to monitor and compare the performance of railway infrastructure managers in the EU.

Arrival airport	Average delay per flight (mins)	Percentage of delayed arrivals
Istanbul-Ataturk	18.4	55.6%
London Gatwick	16.5	45.4%
London Heathrow	13.1	43.1%
Rome Fiumicino	12.0	38.0%
Dublin	12	42.2%
Barcelona	11.9	37.6 %
Lisbon	11.2	39.5%
Brussels National	11.0	41.1 %
Dusseldorf	10.1	38 %
Madrid Barajas	9.6	35.3 %

Table 2: The Top 10 Arrival Airports Affected 2015. Source: Eurocontrol (2016)

According to the RMMS survey in its last report in June 2014 [2], dissatisfaction with punctuality and reliability is highest in France (47%), Germany (42%) and Italy (38%).

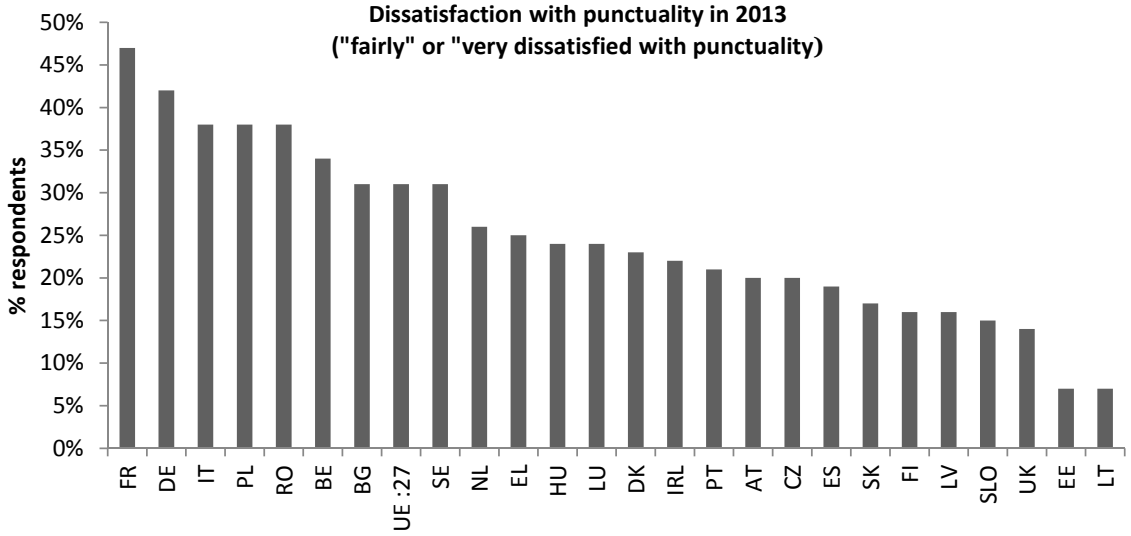

Figure 1: Dissatisfaction with punctuality in 2013. Source: Flash Eurobarometer 382a on Europeans' satisfaction with rail services

This report underlines that there are interesting contrasts in punctuality rates. In Sweden and Italy, long-distance trains have been very punctual in contrast to local trains. However, in Portugal and Lithuania, the opposite has been the case. None of the punctuality rates appear to explain the high degree of dissatisfaction with punctuality and

[2]The Fifth RMMS report covering data up to 2014, should have been published in spring 2016, but it was still not available in October 2016

reliability in France. Maybe this reveals different preferences between European countries, or dissatisfaction due to other service quality components like train cancellations, and which are not considered in the punctuality indicators. Finally, as far as high-speed services are concerned, AVEs in Spain have reached a punctuality rate of 99.2%, whereas in the more congested networks of France, TGVs have reached a rate of 91% (and 85% for the Thalys services in Belgium).

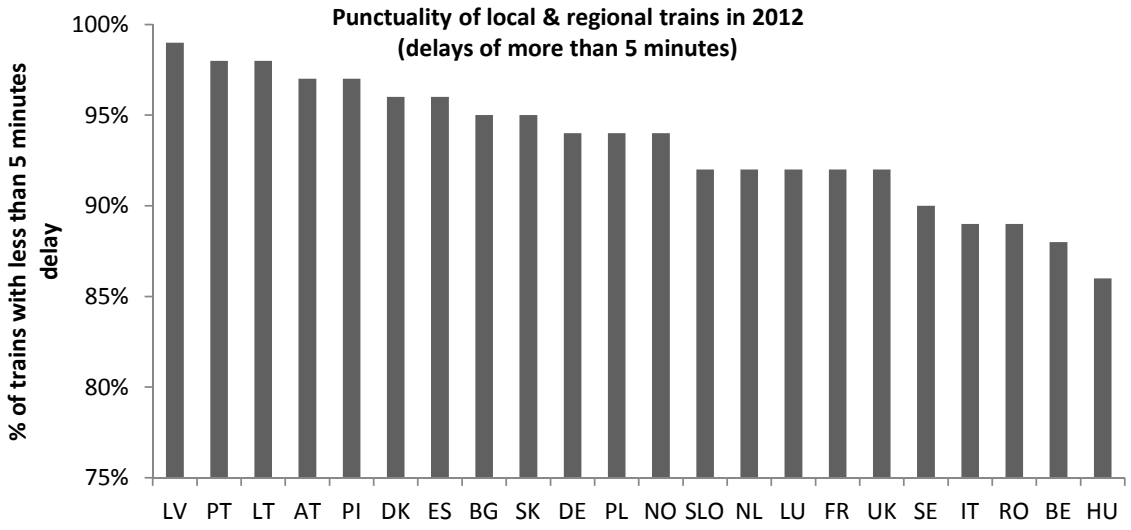

Figure 2: Punctuality of local and regional trains in 2012. Sources: RMMS questionnaires and Trafikverket for Sweden

These figures clearly illustrate the magnitude of the problem of transports delay today.

In many networks, the degradation of service quality is closely related to their degree of capacity utilisation. Indeed, many transport networks suffer from peak load demand in several localised areas, reflecting the costs of capacity constraints.

For example, as stated by the European Commission report "Impact assessment of revisions to Regulation 95/93 (Gleave, 2011), major European airports are facing a capacity crunch, with demand exceeding capacity at some points during the day. Today, five major airports (London Heathrow, London Gatwick, Frankfurt, Paris Orly and Düsseldorf) are considered saturated and operating at full capacity. Capacity constraints at Düsseldorf and Paris Orly are due to policy restrictions (annual slot limit) and not to the real physical capacity of the infrastructure.

Moreover, the projections made by the study estimate that in 2030, 19 main European airports will be saturated including, for example, Paris CDG, Warsaw, Athens, Vienna

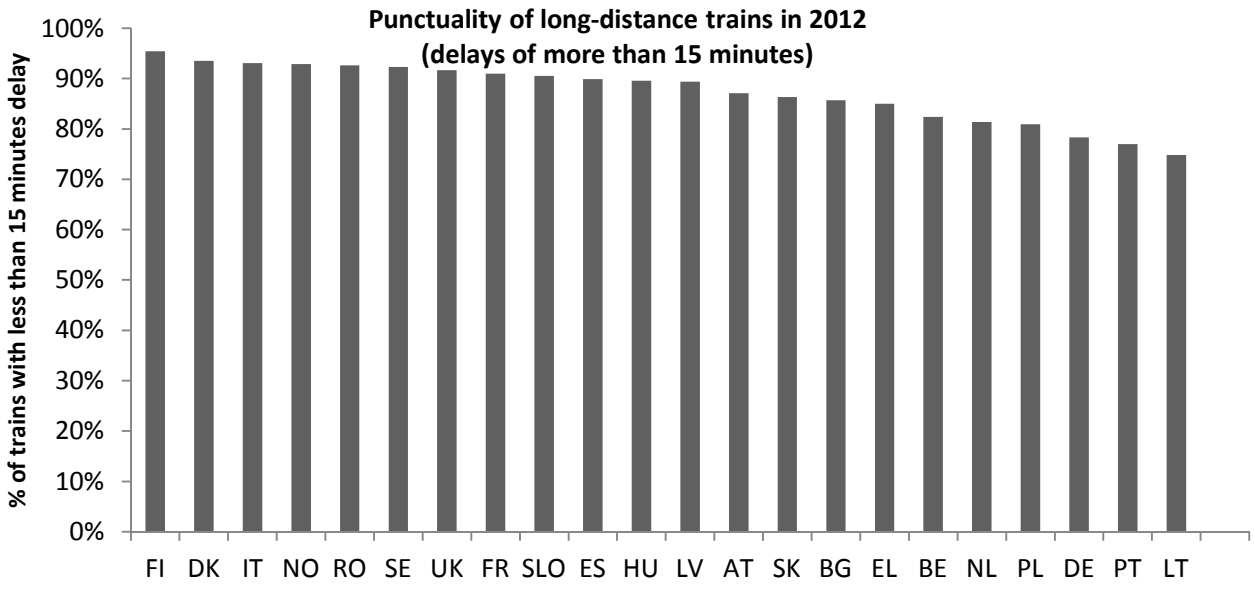

Figure 3: Punctuality of long distance train in 2012. Sources: RMMS questionnaires and Trafikverket for Sweden

and Barcelona. For some airports (those for which data was available) the report also estimates how many hours per day demand exceeds effective capacity.

Airport	2010	2012	2017	2025
Dublin	1	3	0	0
London Gatwick	14	14	14	17
London Heathrow	15	15	15	15
Madrid Barajas	6	12	6	12
Paris CDG	8	11	12	15
Palma de Mallorca	2	2	2	3
Rome Fiumicino	5	6	6	9
Vienna	5	5	9	5

Table 3: Hours per day demand exceeds capacity. Source: Gleave (2011). Note: Covers daytime period (16-18 hours depending on airport).

In the railway sector, the conclusions of the "Mobility 21" commission (Duron, 2013) recommended setting up an observatory for each major railway project justified by saturation issues (e.g. POCL [Paris Orléans Clermont-Ferrand Lyon], Montpellier-Perpignan). The objective is to monitor how the capacity of these lines is evolving and determine if there is effectively a saturation issue.

Technical and Empirical Characterisation of Transport Congestion

The relationship between intensity of usage and service quality is commonly known as **congestion** reflecting the existence of limited capacity on networks for which demand varies periodically.

Arnott and Kraus (2008)proposed a general and contemporary definition of congestion in the New Palgrave Dictionary of Economics:

"Congestion' is the phenomenon whereby the quality of service provided by a congestible facility degrades as its aggregate usage increases, when its capacity is held fixed" .

Considering this definition, congestion is omnipresent in many networks: *"more telephone usage increases the probability of encountering a busy line; higher electricity demand may lead to voltage fluctuations, brownouts and eventually blackouts; more swimmers in a pool make comfortable swimming more difficult; more patients visiting a medical clinic results in longer waits and lower-quality care; in a more crowded classroom, students receive less individual attention, and more time is wasted on administration and discipline; and so on".*

Transport is a service whose quality depends on traffic (Lévy-Lambert, 1968) or subject to overloads (Kolm, 1968). To better understand the specificities of transports, it is interesting to examine in depth how the positive relationship between capacity utilisation and service quality in terms of regularity has been characterised from the technical and empirical perspective in transport:

Road congestion

In road transport, it is well-known that a large number of road users are subject to longer travel times due to traffic jams. As a result, travellers and shippers are confronted by additional travel time, extra costs from wasted fuel and lost productivity.

The standard static model of road congestion is based on "fundamental diagram of traffic congestion" well-known to engineers. This specification describes the speed-flow equilibrium relationships under stationary states initiated by Greenshields et al. (1935) and since improved by significant advances in the 1950s and 60s[3].

[3]For a more detailed description of the speed-flow literature, the reader can referred to (Li, 2008).

The "fundamental diagram" describes that road congestion can be defined by three microscopic fundamental variables: speed S (km/h), flow F (number of vehicles/h) and traffic density D (number of vehicles/km). This relationship is formalised in the form $D = F/S$.

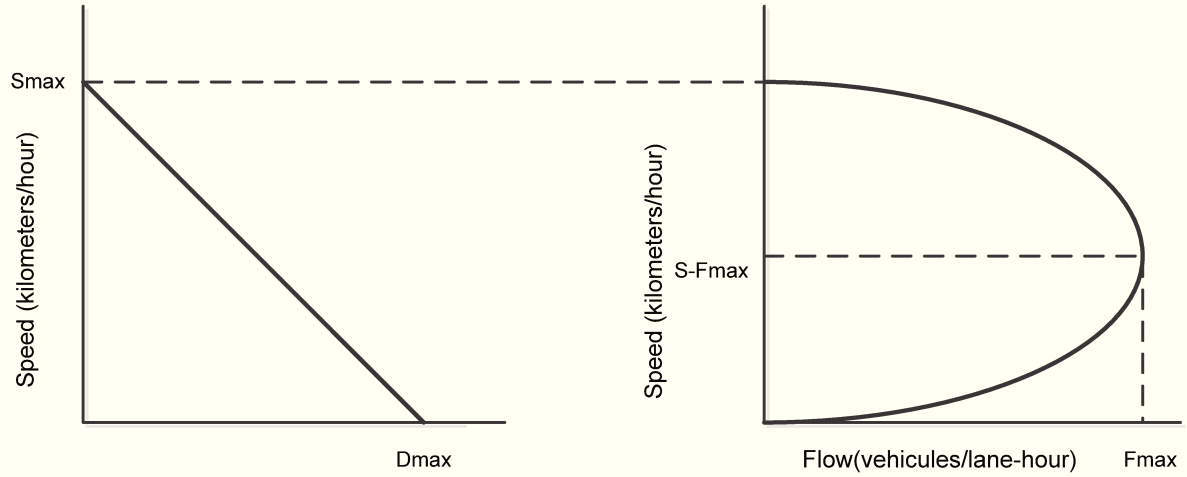

Figure 4: Derivation of the speed-flow curve. Source: Morrison (1986).

Empirical evidence shows that speed S falls with traffic density D. Once capacity has been reached, the addition of a vehicle into the traffic stream causes the flow of vehicles F to decrease, as seen graphically in the second part of Fig.4.

Static speed–traffic flow curves plotted by traffic engineers, which can be observed through road statistics, confirms for roads that a clear relationship exists between infrastructure usage (flow F) and service quality (speed S in this particular case).

Nevertheless, as Quinet (1997) pointed out, congestion does not appear only on roads but also in other transportation modes, even where traffic is scheduled in advance. However, empirical data and literature on congestion is less extensive in this sector.

Air traffic congestion

Much literature has been devoted to airport congestion associated with large airports due to runway or traffic control saturation.

In a seminal paper, Carlin and Park (1970) estimated the marginal cost of delays in New York's La Guardia airport. The congestion cost is defined as the additional delay imposed on the following planes in the queue during the busy period. Further developments proposed to assess the cost of congestion empirically. This is notably the

case of Morrison et al. (1989). Their paper aimed at proving an econometric estimation of the relationship between airport activity and arrival and departure delays using US data. It clearly exhibited that an increasing level of activity causes an increase in average delays. In other words, when capacity is used to its fullest, an additional slot increases the probability of delays due to a reduction in the ability to recover from an incident. Another interesting contribution to this literature was provided by De Rus and Román (2006), who proposed a desegregated analysis of airport delays in Madrid Barajas.

Rail congestion

A considerable amount of literature has dealt with analytical and simulation-based methods in order to study delays and capacity assessment in railroad line haulage networks with specific configurations.

Frank (1966) studied delays on a single track rail line with unidirectional and bidirectional traffic. The author estimated the number of trains that could travel on the network by considering only one train on each link between sidings using single train speeds, and assuming deterministic travel times. This work was later extended by Petersen (1974) to accommodate for two different train speeds, while assuming independent and uniformly distributed departure times, equally spaced sidings and a constant delay for each encounter between two trains.

More recently, Chen and Harker (1990) extended this model to calculate delays for different types of trains over a specified single track section as a function of train schedules and dispatching policies. They assumed a constant probability of delay between trains. Higgins and Kozan (1998) presented a model of urban networks and quantified the expected delays for passenger trains on a complex multitrack rail network. This paper also investigated the influence of modifying scheduled slack time on expected delays. It suggested that, although large reductions in expected delays are achievable with a small amount of slack time, on slight improvements are observed when slack time is increased further (e. g. from 8% to 16%).

Dessouky and Leachman (1995) used a simulation modelling methodology to analyse the capacity of tracks and delay to trains in a complex rail network. Krueger (1999) used simulation to develop a regression model to define the relationship between train delay and traffic volume. Yuan and Hansen (2007) proposed probability models that provide an estimate of delays and the use of track capacity. Murali et al. (2010) presented a simulation-based technique to generate delay estimates over track segments as a function

of traffic conditions, as well as network topology to facilitate routing and scheduling freight trains.

From the empirical point of view, the relationship between traffic density and unexpected delays is quite familiar in airport studies: a primary incident (failure of the rolling stock, failure of the infrastructure, inadequate behaviour of the crew, etc.) can generate delays to the following trains. Given the complexity of the system, a lot of trains can be affected, even on different sections of the network. The transmission of the delay increases as capacity utilisation grows, because heavy traffic reduces the network manager's ability to resolve the incident, thus the delay is transmitted with snowball effect. These kinds of delays are not internalised by the infrastructure manager and can be estimated considering an econometrical approach

Railway delays can be measured with an adequate monitoring system. Very few papers have studied this phenomenon in the economic literature. For instance, it has been studied empirically by the British rail network (Gibson et al., 2002). In this paper, a regression analysis confirmed the existence of a relationship between capacity utilisation and delays. Also, an exponential form was chosen to estimate the relationship between capacity utilisation C_{it} and reactionary delay D_{it} across the network.

$$D_{it} = A_i * exp(\beta C_{it}) \tag{1}$$

The results of the regression show that β is statistically significant for 20 out of the 24 routes. It means that there is a positive relationship between capacity and reactionary delays. This relationship justifies the congestion charge implemented since 2001 by Network Rail. An additional path increases the probability of delays and, therefore, its monetary cost in a performance regime framework. Recently, (Haith et al., 2014) proposed an alternative methodology for the British infrastructure manager which concluded that performance is as much to do with how capacity is utilised as to how much. In other words timetable heterogeneity is an important factor.

Similarly, an extensive econometric analysis has been conducted for the French railway network, with comparable results (Pérez Herrero et al., 2014). This study focuses on 42 lines of the French railway network, with 3 measurement points for each line.

Pérez Herrero et al. (2014), proposed a mathematical framework to estimate the marginal congestion cost of railways empirically. This mathematical framework enables calculating the marginal effect of a train on the total delays.

We define R_i^* as the deviation between the real time and the scheduled time of a train for a given traffic density Q_i. The train can be on time ($R_i^* = 0$), arrive early ($R_i^* < 0$), or late ($R_i^* > 0$).

We define the variable R_i representing the delay of train. It can there be expressed as:

$$R_i = \begin{cases} 0 & \text{si } \mathrm{R}_i^* \leq 0 \quad (2) \\ R_i & \text{si } \mathrm{R}_i^* > 0 \quad (3) \end{cases}$$

The expected delay of train for a given traffic density is:

$$E(R_i) = p(R_i^* \leq 0)E(R_i \mid R_i^* \leq 0) + p(R_i^* > 0)E(R_i \mid R_i^* > 0) \quad (4)$$

As the expected delay is null when the train is on time or early ($p(R_i^* \leq 0)E(R_i \mid R_i^* \leq 0) = 0$), this equation can be written as:

$$E(R_i) = p(R_i^* > 0)E(R_i \mid R_i^* > 0) \quad (5)$$

This equation indicates that the expected delay of a train for a given traffic density is equal to the product of the expected delay of delayed trains and the number of trains delayed. The total amount of delays of trains for a given volume of traffic is, by definition, the expected delay of train multiplied by the number of trains, i.e. $Q_i E(R_i)$. Therefore, it follows that the marginal delay imposed by an additional train is the derivative of the total amount of the delay function relating to the level of traffic.

It can also be written as:

$$\frac{\partial Q_i E(R_i)}{\partial Q_i} = Q_i \frac{\partial E(R_i)}{\partial Q_i} + E(R_i) \quad (6)$$

In this equation, the second right hand term is the expected delay of the additional train given the traffic density: this is a direct effect internalised by the train. The direct effect is equal to the expected delays for a given traffic density. This term, expressed by equation 5, can be directly computed from the data set.

The first right hand term of equation 6 represents the marginal delay imposed by the additional train on the following trains. It is an indirect effect which corresponds to the pure externality effect of congestion. The indirect effect, cannot be computed directly

and needs and econometric analysis in order to be estimated[4].

According to the line and its features (allowed speed, number of tracks, signalling), the results show a positive econometric relationship between the traffic and the unreliability rate or the length of delay: an additional train on the line increases the probability of delays, for itself and for the other trains. The marginal congestion cost is made up of a direct effect which is internalised by the supplementary train and of an indirect effect that generates an external cost on next users.

Towards a Microeconomic Approach of Rail Capacity Constraints

Econometric results confirm that a supplementary train increases delays and therefore travel time costs for users. Beyond a certain traffic density threshold, additional production (train path) would increase per-unit costs (for users and train operating companies) in the railway system leading to decreasing returns from density, at localised times and in areas. This finding could seem surprising in the railway sector, a network commonly associated with increasing returns to scale.

Rail transport, like other public utilities, has been traditionally considered as a natural monopoly, describing a market in which for structural reasons, it is more profitable than if just one firm produces a service. The concept of natural monopoly was initially applied in the literature by Mill (1848) and Dupuit (1854)[5]. Until the end of the 1970s, the definition of natural monopoly was closely related to the concept of economies of scale. As Samuelson (1948) stated:

"Some of the basic factors responsible for monopoly are inherent in the economies of large-scale production".

It was considered that in some activities, technology involves very high fixed costs, such as creating and maintaining rail infrastructure and services for example, and very small marginal costs for providing an extra unit. Once the infrastructure and train equipment are determined, it costs very little to increase an extra unit of rail traffic and implies a declining average cost curve. In fact, the firm's average cost decreases as input increases, because the fixed costs are shared between a greater numbers of output units.

[4]A detailed analysis of the data base and the econometric results are proposed in Appendix A

[5]For a more detailed literature review on the concept of natural monopoly, the reader can refer to (Mosca, 2008)

As Mosca (2008) described in an article which analyses down to the last detail the origin of the concept of natural monopoly, the traditional definition of natural monopoly was criticised at the end of the 1970s. At this period, Baumol et al. (1982) considered that the concept of economies of scale was not a sufficient condition to define a natural monopoly. It became apparent that it was the concept of subadditivity [6] rather than the degree of scale economies which defines the concept of natural monopoly. This precursor paper of Baumol et al. (1982) focused in the case of a natural monopoly with multi-product firms by introducing the theory of contestable markets.

At the end of the 1980s, the concept of contestable market was criticised in turn, since in some cases *"market forces are unlikely to reduce market power in a number of cases (if sunk costs are high, if consumers have switching costs, if there are network externalities and if monopolists can engage in anti-competitive market practices"* (Motta, 2004).

Even if the concept of economies of scale is clearly not the only relevant attribute for defining a natural monopoly theoretically and mathematically, it has been traditionally associated with its definition.

Network industries such as railways, telecommunications and electricity were considered natural monopolies[7] until the end of the 20th century and thus they have been naturally linked to the concept of scale economies.

At the beginning of the 1990s, in the context of railway deregulation, there was increasing interest in the definition and measure of economies of scale in this sector [8]. The studies by Caves et al. (1981) and Caves et al. (1984) were the first to estimate whether American rail companies presented scale economies. These estimations were also carried out in some European countries by Preston and Nash (1996) Cantos and Maudos (2001) and Cantos (2001).

Although the aim of these studies was to measure rail system efficiency, they also improved knowledge of the relationship between operating costs and railway infrastructure production (train paths). In this period, specific theoretical developments concerning the cost function for transports firms, recommended using two indices to better analyse the structure of the transport industry: returns of density (RTD) and returns of scale with

[6]The subadditivity condition implies that production from only one firm is socially less expensive (in terms of average costs) than production of a fraction of the original quantity by an equal number of firms.

[7]Nevertheless, Perennes (2014) recalled that the French rail sector was not organised as a monopoly until 1937.

[8]For a complete and exhaustive review on the testing for economies of scale in rail transport, the reader can refer to Oum et al. (1999).

variable network size (RTS) (Caves et al., 1984). The first, RTD, refers to the impact of expanding traffic on average cost, but holding network size constant while the second, RTS, measures the impact of a proportional increase in traffic and network size on average cost (Oum et al., 1999).

Since then, there have been numerous empirical calculations of RTD and RTS in the literature for different industries. As summarised by Basso et al. (2011), studies such as that of Braeutigam (1999) for railways showed that there were increasing returns to density (which means that it would be advantageous for industries to increase traffic density on their networks), but constant returns to scale (meaning that there is no clear empirical evidence of the cost advantage of expanding networks).

Nevertheless, present mobility patterns with traffic concentrations in a small number of cities and the new theoretical economic developments described in this introduction, now raise the question of rail capacity constraints and decreasing returns to density in the rail sector. Stated differently, we can observe that from the social angle (considering user's costs), the railway industry can be characterised in some places and a precise moments by decreasing returns to density: a conurbation leads to congestion.

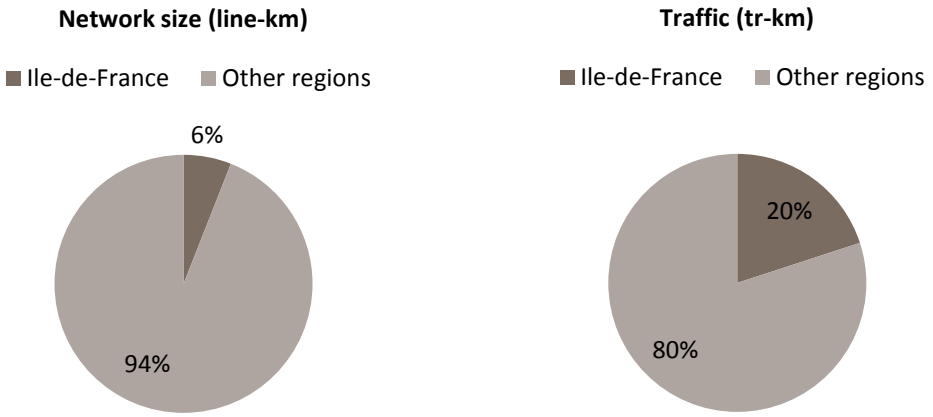

Figure 5: Network size and traffic repartition for Ile de France. Source: SNCF Réseau

Populations and by consequence travellers, are now concentrated in determined metropolitan areas in which, furthermore, the investment cost of expanding capacity is particularly high. As shown in Figure 5 the Parisian metropolitan area concentrates 20% of total rail traffic but only represents 6% of the network size. These figures illustrates that there are some areas in France with very high traffic density levels, but also that this

phenomenon is not constant around the network and, in some cases, it is also temporally localised.

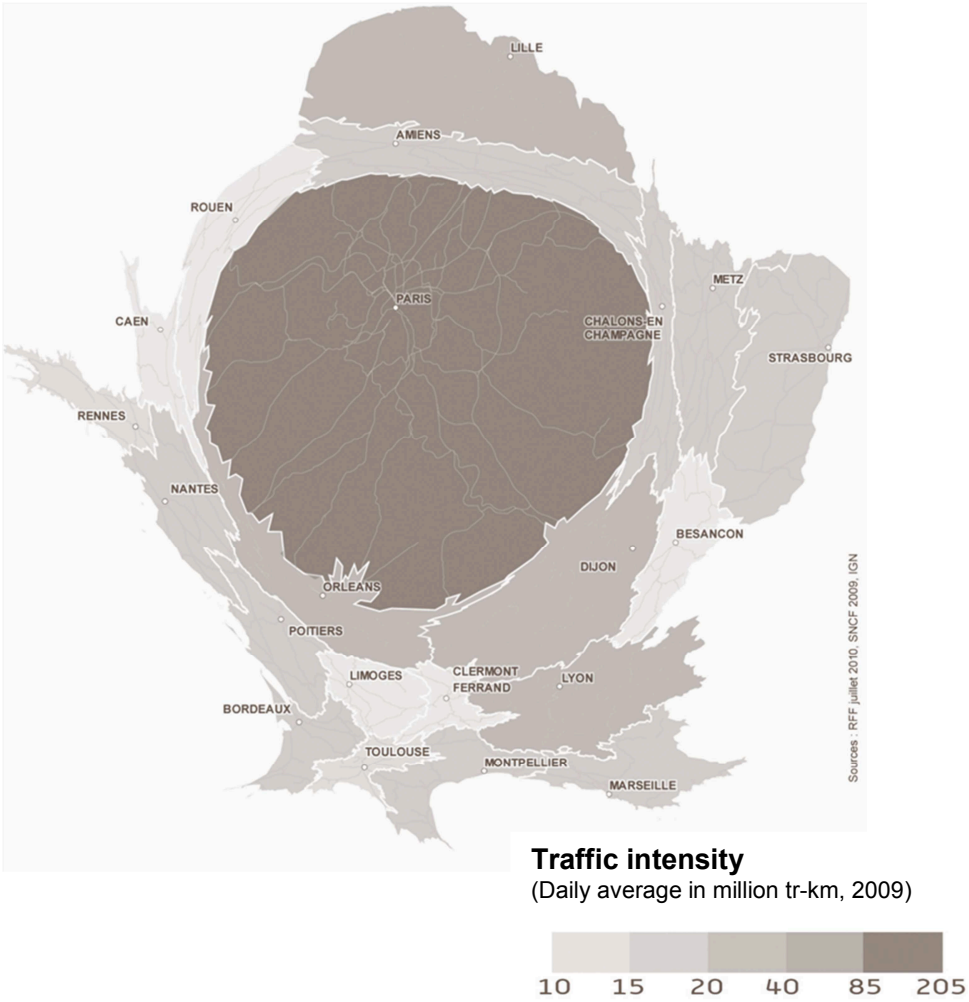

Figure 6: Per region traffic intensity. Source: RFF 2010, SNCF 2009, IGN

Content of the Dissertation

This introduction has illustrated that the relationship between service quality and available capacity is a major challenge for the railway sector.

Moreover, the economic developments that have occurred during the past few decades have provided better understanding of the market structure and improved cost measures for network industries, particularly for the railway sector. In this context, this PhD dissertation aims to analyse in detail the reasons for and consequences of an increasing average cost associated with an increment of traffic during peak hours at localised areas of the network. Put another way, it studies in detail the economic characterisation of decreasing returns to density in a very busy rail network area.

The structure of the PhD dissertation is as follows. chapter 1 proposes a review of the engineering definition of rail capacity and its main components. This description allows demonstrating that service quality is a key parameter in the definition of capacity. This chapter details precisely the operational and empirical issues at stake when facing problems of capacity constraint in railway transportation. Furthermore, it describes and compares the differences in the process of building the graphic timetable in some European countries and identifies the different methods employed in each network for ensuring robustness.

After having understood the scope and limits of the engineering perspective, chapter 2 develops the theoretical economic framework of transport congestion. The purpose of this literature review is to identify how congestion has been studied from an economic perspective in other sectors and understand the similarities and differences with the rail sector.

Considering the engineering specificities of the rail sector described in chapter 1 and the economic lessons from chapter 2, chapter 3 proposes a new generalised user cost function approach. The microeconomic model described in this chapter incorporates a theoretical measure of the value of frequency for transport services with timetables. This model identifies and formalizes mathematically the consequences for users of having different rail frequencies in terms of service provided (expected schedule delay cost) and delay (random delay cost).

Once the detailed generalised cost function for train users has been determined, chapter 4 seeks to build an equilibrium model, by considering users' behaviour and operators' costs, and describing how supply and demand interact under different market conditions.

The aim of this chapter is to analyse the interactions between demand and supply, and discuss if standard theoretical conclusions in other transport sectors, such as price mechanisms, are optimal tools for dealing with congestion, and under which conditions, taking into account the specificities of the rail sector developed in the previous chapters.

The last part of this dissertation "Conclusions and policy recommendations" point out the main issues of the research dealt with in chapters 3 and 4. In order to propose certain practical policy implementations, the theoretical recommendations made in the previous chapters must be considered in the light of the legal European network. We observe that even if the regulatory context allows infrastructure managers to identify and declare capacity constraints and to price them under certain conditions, few European countries apply this procedure. In practice, reflection on rail capacity constraints and the possibility of implementing a congestion price now gives food for thought in infrastructure managers' procedures, but it is still at a preliminary stage. In this context it seems important to propose further policy recommendations that will stimulate and develop the economic debate in the next few years.

Chapter 1

The Notion of Railway Capacity: Definition and Key Parameters

Contents

1.1 Introduction

In order to use the railway infrastructure, train operators companies must reserve part of its capacity. Insofar as the capacity of the railway infrastructure is fixed in the short term, more intensive use of this infrastructure can have negative consequences for its users, if such use affects the quality of the service.

The measure and estimation of capacity constraints thus plays an essential role in all decisions related to the attribution of capacities and investment. Nonetheless, current methods of analysis do not allow integrating the value of capacity constraints in the infrastructure manager's decision making process.

In addition to the observatories recommended in the conclusions of the "Mobility 21", a scientific committee has been set up in order to define a measure of infrastructure saturation since it is a difficult if not controversial subject. Likewise, with the methodological framework currently used for socioeconomic evaluation, it is difficult to estimate the advantage for society of creating capacity for a development project.

In the European Union's legislation, the regulatory framework of the railway capacity attribution and pricing process is set out in chapter IV of Directive 2012/34/UE called *"Pricing of railway infrastructure and distributions of railway infrastructure capacities"*. The declaration of saturation of a railway infrastructure and the actions to be carried out are defined in the directive as follows: *"Where, after coordination of the requested train paths and consultation with applicants, it is not possible to satisfy requests for infrastructure capacity adequately, the infrastructure manager shall immediately declare that section of infrastructure on which this has occurred to be congested. This shall also be done for infrastructure which can be expected to suffer from insufficient capacity in the near future"*.

The notion of capacity constraint as defined in the directive refers to capacity allocation processes (*"to satisfy requests for infrastructure capacity adequately"*). It partially expresses the incapacity of the network to accommodate flows. The works of IRG rail complete this vision by considering the difficulty of the infrastructure manager to allocate train paths without diminishing the quality of service. However, being able to identify the level of flows above which one speaks of capacity constraint requires defining the capacity of an infrastructure beforehand.

There is no general consensus on what capacity constraint is in the railway sector. A precise definition of capacity and the elements that compose it is required to clearly identify the levers of action to overcome capacity constraints. The purpose of chapter

1 is to examine the concept of capacity and its components in detail from the angle of engineering. Traditionally, the railway system and its technical functioning have been defined and analysed by engineers. In view to providing a complete economic analysis of railway capacity constraints, it is important to have precise understanding of the definition of capacity and especially of the rail timetable process seen from the standpoint of engineering.

1.2 The Concept of Capacity

Capacity is an essential notion for railway infrastructure managers (IM). However, although the definition of "capacity" is frequently used, this term is complex and has no genuine definition or standardised measure.

1.2.1 Capacity according to different viewpoints

In the railway industry, capacity can be defined as a maximum volume of traffic. From the viewpoint of final demand for transport, it has been defined by some as follows: "*Capacity is a measure of the ability to move a specific amount of traffic over a defined rail line with a given set of resources under a specific service plan*" (Krueger, 1999).

This definition stands for the final quantity transported. It considers without distinction the capacity of the infrastructure and that of train operating companies (TOC). Although this definition is generally used to express rail transport capacity in relation to other modes of transport; it is rarely used in daily railway operations.

In the practice of network management, railway capacity can be considered as:

- *"The maximum number of trains that can traverse the entire railway or certain critical (bottleneck) section(s) in a given duration of time"* (Burdett and Kozan, 2006).

- *"The highest volume (trains per day) that can be moved over a subdivision (plant) under a specified schedule and operating plan (traffic and operations) while not exceeding a defined threshold (over-the-road-time)"* (Krueger, 1999).

- *"The capacity of any railway infrastructure is*:

– *The total number of possible paths in a defined time window, considering the actual path mix or known developments respectively and the Infrastructure Manager's own assumptions;*

– *in nodes, individual lines or part of the network*

– *with market-oriented quality"* (Union Internationale de Chemins de Fer, 2004).

Market (customer needs)	Infrastructure planning	Timetable planning	Operations
expected number of train paths (peak)	expected number of train paths (average)	requested number of train paths	actual number of trains
expected mix of traffic and speed (peak)	expected mix of traffic and speed (average)	requested mix of traffic and speed	actual mix of traffic and speed
infrastructure quality need	expected conditions of infrastructure	existing conditions of infrastructure	actual conditions of infrastructure
journey times as short as possible	time supplements for expected disruptions	time supplements for expected disruptions	delays caused by operational disruptions
translation of all short and long-term market-induced demands to reach optimised load	maintenance strategies	time supplements for maintenance	delays caused by track works
		connecting services in stations	delays caused by missed connections
		requests out of regular interval timetables (system times, train stops, ...)	additional capacity by time supplements not needed

Figure 1.1: Different approaches to capacity. Source: Union Internationale de Chemins de Fer (2004)

More generally, capacity is defined in several ways according to context and need. Figure 1.1 compiles the different viewpoints on the term "capacity": market, infrastructure planning, timetable planning and operation.

From the market standpoint, demand for capacity is oriented to the satisfaction of passenger demand at peak hours whereas infrastructure planning, on the contrary, tends to define capacity, which on average, guarantees optimal utilisation of the network.

From the viewpoint of timetable planning, measuring capacity has to take into account the type and characteristics of the infrastructure as well as existing demands for train paths. A strong link has been demonstrated between capacity and the differences of average speeds between trains, a difference in turn linked to service policies as much as to pure speed. Lastly, from the operational standpoint, the definition of capacity depends on

the availability of the infrastructure and must take into account its real traffic conditions, the number of trains and incidents at a given moment. If each definition is relevant according to the context in which it is used, this leads to different calculations of capacity need.

Following the conclusions of the "Mobility 21" commission, a scientific committee has designed a pedagogical kit to define a common measure of infrastructure saturation. Following its recommendation, the Montpellier-Perpignan observatory, which monitors how the capacities of these lines are evolving, has recently published its final report (Rebeyrotte, 2016).

1.2.2 Capacity parameters

The elements presented above reveal that the measure of railway capacity is not absolute. It greatly depends on the way it is used. According to Abril et al. (2008), the main determinants of capacity can be classified into three categories:

- Infrastructure parameters
  - Signalling block system
  - Single line/Double line
  - Network effect
  - Structure of the line and speed limits
  - Block length
- Traffic parameters
  - New and existing lines
  - Traffic mix
  - Clock-face timetable
  - Distribution of traffic: peak hours, off-peak hours
  - Priority rules
- Operational parameters
  - Line interruptions

- Station stopping time
- Maximum journey time
- Time unit
- Service quality: reliability, service robustness

According to this classification, the quality of service is one of the main factors to consider when evaluating the network capacity of an infrastructure and a transport plan in the short term.

1.2.3 The role of service quality requested in the definition of capacity

Different definitions of capacity are generally used when considering the impact of quality of service in the railway sector (Krueger, 1999):

- *Theoretical capacity*: this corresponds to the number of trains that can run on a line during a determined time interval, in a mathematically perfect environment with trains circulating continuously with minimum headway (time interval between two consecutive trains). This measure is the ceiling of the line's capacity. Most usually, homogenous traffic is assumed with the same rolling stock and circulations distributed throughout the day without disturbances. This measure therefore omits the heterogeneity of traffic and commercial transport plans (diversity of stop, speed policies, etc.). Furthermore, it does not take into account any buffer time in the graphic timetable. The theoretical capacity can be calculated by using a theoretical formula.

- *Practical capacity:* this corresponds to the practical limit of a volume of traffic considered as "representative" on a line, with a predetermined level of reliability. It is measured by the number of trains that can run per unit of time with a level of operating quality statistically equal to the level desired (excluding major incidents). The representative traffic reflects the current combination of trains (transport plan), priority rules, etc. If the theoretical capacity represents the line's upper limit of capacity, in theory the practical capacity represents a measurement unit calculated on the basis of hypotheses taking into account the hazards occurring during operation.

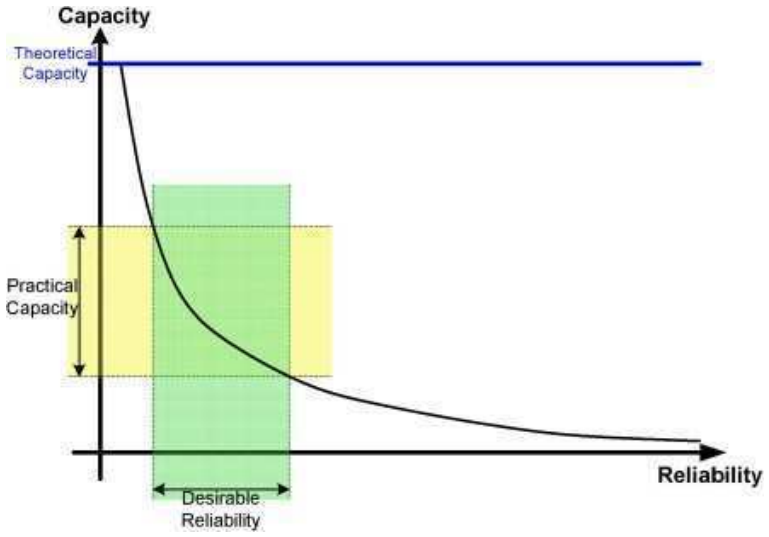

Figure 1.2: Practical capacity involves a desired reliability level. Source: Abril et al. (2008)

- *Used capacity*: this is the current volume of traffic on a line or in the network. It is lower than or equal to the practical capacity as a function of the type of line analysed.

- *Available/residual capacity*: this is the difference between the capacity used and the practical capacity. It indicates the volume of additional traffic that could be accommodated by a line.

The choice of service quality in a transport plan is closely related to the definition of practical capacity. All things being equal, the desire for a high level of reliability results in a lower practical capacity. For railways, as well as for other modes of transport, there is a statistical relation between traffic density (thus used capacity) and service reliability/quality (Pérez Herrero et al., 2014).

The literature (Krueger, 1999; Abril et al., 2008) includes different recommendations on the levels of line use that allow defining practical capacity. For example, the UIC 406 code (Union Internationale de Chemins de Fer, 2004) recommends setting the maximum rate of use of a high speed line at 75% of its theoretical capacity at peak-hours, and at 60% of its theoretical capacity as the daily average.

According to the criterion of the UIC 406 code, the recommendations of these maximum rates of use correspond to a reasonable level of reliability for a given infrastructure. If this level is exceeded, it entails a risk of degrading service quality. Despite the use

Type of line	Peak-hour periods	24 hours
Dedicated to suburban commuter traffic	85%	70%
Dedicated to high speed trains	75%	60%
Mixed traffic lines	75%	60%

Table 1.1: Proposed limits of occupancy rates. Source: UIC 406 code

of the UIC 406 code, using maximum rates of occupancy as a criterion of network reliability, and the fact they have been defined empirically, the sheet does not specify the procedures for calculating the recommended maximum rates of use . In practice, rates of occupancy can exceed the UIC's thresholds without deteriorating operating quality. Conversely, poor operating quality does not necessarily imply rates of occupancy higher than the UIC maximum.

1.3 Methods of Railway Capacity Evaluation

As stated above, different methods are used in order to evaluate railway capacity as a function of the different definitions of capacity, the precision of the available data and the need for detail in the estimations.

As described by Abril et al. (2008), the most significant methods can be classified into three levels: analytical, optimisation and simulation.

1.3.1 Analytical methods

These are simple models whose objective is to determine a preliminary solution, giving major indications on the level of utilisation of an infrastructure. These methods have been obtained through mathematical formulations. They enable defining reference values regarding line capacities or comparisons between lines. These calculations are used to determine the theoretical and the practical capacity as a percentage of the former.

An example of using these methods was given in a work of Petersen (1974). In his article he aimed at measuring capacity using an analytical model with a single line with uniformly distributed departure times and three types of different train speed. In the 1990s, existing analytical models were completed by the works of Chen and Harker (1990) and Harker and Hong (1994). Those studies proposed to estimate delays linked to a certain traffic density by using a stochastic approach for a single and double line, respectively. More recent works on analytical estimations including those of Burdett and Kozan (2006),

permits estimating theoretical capacity by varying a large number of parameters (train heterogeneity, train spacing, localisation of branch lines, etc.).

As mentioned by Abril et al. (2008) and by Kontaxi and Ricci (2010), these methods propose useful results for identifying major capacity constraints and are relatively easy to obtain. However, their main disadvantage is that the results vary considerably from one method to another and they are very sensitive to the parameters used.

1.3.2 Optimisation methods

These methods provide more accurate estimations and strategic solutions for capacity problems than analytical formulas. Optimisation models are mainly based on obtaining saturated optimal times using programming techniques (e.g. Mixed Integer Linear Programming Formulations and Enumerative algorithms). The optimisation method based on saturation obtains the capacity of a line by programming a maximum number of additional trains for a predetermined schedule.

This method has led to the establishment of a sheet by the Union Internationale des Chemins de Fer (UIC 406 code) intended for infrastructure managers. The UIC 406 code sets out a method of timetable compression by reducing the headway time between trains, while conforming to the minimum time necessary to clear a line. The time remaining after compressing traffic theoretically represents the time available for additional traffic.

For more details on optimisation methods, Abril et al. (2008)released a highly detailed document with technical review of these methodologies and their developments.

1.3.3 Simulation methods

Simulation methods represent the most sophisticated and detailed stage of measuring railway capacity. They permit the reproduction of reality and railway operating processes by using software applications to evaluate capacity following changes in the transport plan and their impact on the robustness of the graphic timetable. In addition to purely theoretical models, some simulation programs have been developed and are available on the market. These simulation programs make it possible to implement theoretical analytical methods at the industrial level. A few examples of these software applications are:

- Open track

- MultiRail
- RailSys
- Rail Traffic Controller

For a more detailed description of simulation models and these software applications, the reader can refer to Barber et al. (2007)),Transportation Research Board (2013) and Hansen and Pachl (2014).

1.3.4 A study on the indicators of railway capacity utilisation at SNCF Réseau

SNCF Réseau, the French IM, launched a study in 2015 on the indicators of the rate of line use. It was carried out in the framework of scientific recommendations formulated by the saturation observatory and in view to focusing on how the real use of a line can be measured and compared to the capacity that it can supply theoretically. This study was performed by the company INGEROP and aimed at highlighting one or more pertinent indicators of the level of use of a line and which are capable of characterising its level of saturation. The final report proposes an analysis and a comparison of different methodologies available in the literature for evaluating the level of line use. Three approaches were implemented:

- *The infrastructure occupancy approach*, by calculating a rate of occupancy. This method refers to the application of the UIC 406 code intended for infrastructure managers. The approach consists in calculating the occupancy time in a graphic timetable by compressing train paths on a section of line and over a predetermined period.

  The method proposed by the UIC code determines that compressing must be done on all the elementary sections of a line defined beforehand by interlocks (junction, crossing rails, etc.). The study performed by INGEROP also proposed compressing an entire commercial line (and not simply each of its elementary sections). The objective of this alternative method was to provide information on the additional available capacity from the commercial standpoint and not simply from that of operating the infrastructure. This method is very similar to the CUI approach (Capacity Utilisation Index) used in the United Kingdom by the infrastructure manager Network Rail in its capacity analyses.

- *Approach by robustness*, through simulation and measuring the resilience of a service/infrastructure couple, i.e. the aptitude of an infrastructure to absorb disturbances. This method is mentioned in appendix 8 of Network Statement for the Service Schedule 2016 (SNCF Réseau, 2016) and in an internal document concerning network robustness SNCF (2001) relating to the robustness of graphic timetables. The robustness approach consists in simulating the consequences of a disturbance on a train at a given moment and point of a line. The usual indicators stemming from this method are the number of trains affected, the time to return to normal and the number of minutes lost locally.

- *Approach by timetable variances* , using a statistical analysis of delays observed for trains. This method is suggested for analysing the times achieved by trains in comparison to a planned timetable in order to determine the propagation of delays and the zones where delays are triggered. The indicators given by the delay analysis approach stand for the distribution of timetable variances, the average delay at a given point and the rate of increase of delays on a section.

1.4 The Current Process of Building a Graphic Railway Timetable in France

All countries and their railway systems have their own structures and methods for building graphic timetables.

From the design stage, the infrastructure manager (IM) incorporates the objective of ensuring the reliability of the train paths supplied. From the manager's standpoint, this objective is integrated in the travel time design and in the train headway distance rules. The rules relating to travel times and additional headway ensures the robustness of the train path and the graphic timetable, respectively. (Verchere and Djellab, 2013).

The following section provides a description of the graphic timetable construction process in France. The aim is to clearly identify the principles and effects of the rules of robustness of the train path and the graphic timetable.

1.4.1 The robustness of travel time and margins of regularity

The timetable of a train is determined on the basis of a travel time calculated as a function of the characteristics of the train and the line travelled as well as commercial and technical

constraints.

According to the internal document "Determination and formulation of timetables" (DCF-DPS Supervision et Support. Réseau Ferré de France, 2006), the travel time of a train path is the sum of four elementary times:

- Basic running movement,
- Stopping time
- Regularity margin,
- Additional time required to build the graphic timetable (traffic halts, extension of stopping time claimed and domestication), and possibly an additional margin for works.

The timetable is defined as follows:

1. A basic running movement that represents the net travel time. This is defined as the result of the calculation that takes into account a given item of rolling stock and a given infrastructure. It is calculated as a function of:
    - The traction unit
    - Gross trailing load
    - The type of rolling stock towed
    - The characteristics of the line travelled (profile, speed limits, power supply, etc.)
2. Commercial stopping times and service stops demanded by railway companies.
3. The rail travel time always includes a regularity margin that allows absorbing part of the delays caused by:
    - traffic production hazards (about 2min/100 km on classical lines)
    - times lost linked to works or maintenance management (about 2.5 min/100 km on classical lines)

    On a classical line, the usual regularity margin is calculated in minutes per 100 km, with a value of 4.5mn/100 km. For high speed lines, the margin is calculated as

a proportion of time and not distance. The normal regularity margin is 5% and exceptionally 7% on the Nord high speed line.

Upon request from the IM, some exceptions can exist in the value of margins on certain classical lines:

- limitation to 3 min/100 km for certain designated passenger trains;
- increase to 5.5 min/100 km for train paths limited to a speed of $\leq$ 100km/h or during certain maintenance time zones notified by the IM.

Furthermore, on the network of Paris region, specific margins are also applied. In this part of the network with dense traffic, the regularity margin (corresponding to 5% of the basic operation) is applied for all trains (DGDI Bureau des Horaires, SNCF, 2007).

4. In some circumstances, the IM can allocate an additional margin (to the regularity margin for works) to offset the time lost generated by specific works. On the contrary, train operating companies (TOC) can also ask for a lower margin than the norm on certain journeys, under their commercial responsibility and in order to offer attractive travel times.

These margins are added to the basic running time. Once the volume of the margins has been determined, they must be distributed. Their efficiency depends on their distribution approach. Table 1.2 presents three different methods of distribution with their associated advantages and disadvantages.

To guarantee a certain level of service quality, the regularity margins added to the basic operation undoubtedly allow offering a reliable railway service with a robust travel time subject to slight variations. However, the travel time of each link is increased systematically in the timetable construction. Thus, we see that the desire to offer travel times with a certain level of reliability, leads to lengthening the travel time of each train trip.

1.4.2 The robustness of the graphic timetable and additional headways

In the previous section, we described the rules applied to the construction of timetables which ensure a certain level of robustness for a given train path. Nevertheless, each train

Distribution of margin	Advantages	Disadvantages	Case of application
Method 1: Linearity	Simplicity of the calculation. Entitlement to hazards over the whole journey.	Small margin at the end of the journey and possible waste at the beginning of the journey	General case (except lines where temporary speed limits impose major slowdowns.
Method 2: Concentration on a section following a zone with predicted or probable loss of time	Lost times are quickly absorbed and do not impact the entire journey downstream.	Small margin available on the other sections.	Work zones with temporary speed limits. Zones with high risks of traffic hazards.
Method 3: Concentration of a large share at the end of the journey or around a major hub of the network	Favours the regularity of the train at its terminal (contractual commitment of the IM). Aids the punctual departure of the train using the same rolling stock reutilised at the terminal.	Small margin available on upstream sections. At the end of the journey the trains are sometimes barely occupied: few customers arrive on time while others, alighting during the travel, are late.	Long distance trains with reutilisation of rolling stock very shortly after their arrival.

Table 1.2: Modes of distributing the regularity margin. Source: Verchere and Djellab (2013)

path is integrated in a graph that includes other traffic. In order to integrate a train path in the graphic timetable, layout standards define the headways that must be introduced between train paths. These standards are defined by types of train and section.

Firstly, the headway between two successive trains depends on a minimum technical value of headway between two trains taking into account signalling systems and safety standards. An additional headway is added to this minimum technical headway to impact on the robustness of the graph. The additional headway can take the following forms:

- A "free block" time. The train cannot cross the signal at the same time as it changes to green. The signal must show beforehand the indication that the line is clear for a minimum time, so that the driver sees the "free block" signal when approaching the panel (and is not tempted to anticipate braking in view to finding a restrictive indication from afar). This time is designated by the Greek letter χ (khi) and is generally equal to 35 seconds (SNCF Réseau, 2016).

- The values calculated are rounded with a precision consistent with the reliability of the train traffic in the zone, generally 30 seconds or one minute (SNCF Réseau, 2016)

- An additional headway (called "buffer time") permits minimising the transmission of delays between trains. Buffer times can take the form of a uniform additional headway between all the train paths or of a "buffer train path" that corresponds to a train path not used and left empty between two consecutive trains.(Normes de tracé horaire en ligne pour le SA, 2015).

As described with the regularity margins, and still with the intention of ensuring a certain service, additional headways are introduced to minimise the transmission of delays in case of an incident. However, additional headways between trains consume capacity and reduce the effective capacity of a line.

1.4.3 Conclusion on the robustness of travel time and the graphic timetable

The methods described in the previous sections show that the current process for building train paths and graphic timetables in France includes a rationale on robustness and the quality of service provided. All things being equal elsewhere, there is a trade-off between the robustness of the train path and travel time, and between the robustness of the graphic timetable and capacity. Nevertheless, the level of these trade-offs does not appear objectivised as yet, but appears to be the result of trial and error or of a standard definition to industrialise the process of building train paths. Figure 1.3 shows the different methods described and used today for graphic timetable construction.

Therefore the regularity margin must not be confused with additional headways. Regularity margins enable trains to catch up a slight delay during their journey, whereas additional headways prevent the transmission of delays between trains. In addition, regularity margins increase train travel time, whereas additional headways reduce the quantity of trains that can be programmed without modifying their travel time.

It is therefore important to bear in mind that the limit of practical capacity is determined by choosing a given level of reliability. The IM therefore has to be able to justify the practical capacity defined for each line as a function of the reliability rules defined beforehand and applied.

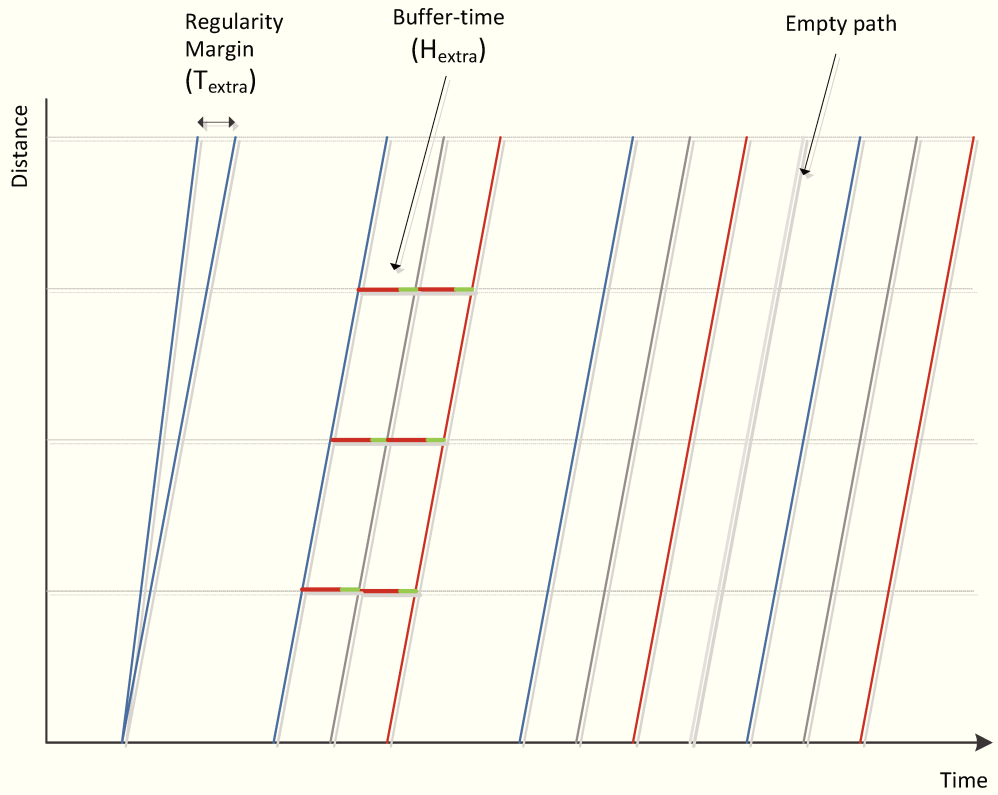

Figure 1.3: 3 Methods for improving robustness in building the graphic timetable: regularity margin, uniform buffer-time and buffer train path. Source: SNCF

In the case of buffer train paths, it appears easy to measure the unused capacity dedicated to controlling robustness. On the contrary, determining the capacity neutralised by using uniform buffer-times between trains still remains a complicated task. The layout standards used for building the graphic timetable define the headway rules to be applied by train schedulers. There is no clarification in these standards between the share corresponding to the gross technical headway, the share corresponding to rounding off and the share accorded to extra buffer-times.

Identifying the reliability standards that define the practical capacity of a line will allow the IM to justify the costs and advantages of changing the current rule. Moreover, if reliability is an end in itself, one may question how and by whom it should be determined.

In the current practice (SNCF Réseau, 2016), basic graphic constructions (train paths with a frequency higher than once a week) are considered robust so long as an isolated disturbance of 10 min on a train is absorbed after one hour from the time of occurrence.

1.5 The Experiences of other European IMs in the Graphic Timetable Construction Process.

In the context of this research it is interesting to analyse experiences of managing capacity constraints in different railway systems, either by country or by type of network. The main idea is to identify and compare the similarities and differences in timetable building processes employed by different infrastructure managers, in order to compare SNCF Réseau's practices with those of other networks.

The present section describes the details and rules of the timetable production processes of several European network managers that answered my requests for the purposes of this research.

1.5.1 Régie Autonome des Transports Parisiens: RATP

RATP is a publicly owned company (public commercial and industrial establishment) that operates public transport in the Ile-de-France region (notably the subway and regional express network). It maintains, upgrades and develops one of the densest multimodal networks in the world: fourteen subway lines in Paris, eight tram lines (T1, T2, T3a, T3b, T5, T6, T7 and T8, with line T4 operated by SNCF), part of the bus lines of Île-de-France, and most of lines A and B of the regional express network of Île-de-France (RER).

RATP fulfils its mission of supplying public transport as part of long-term operating contracts signed with the Transport Syndicate of Île-de-France (STIF), which is the transport organisation authority of Île-de-France.

With more than 1.7 billion passengers carried every year (source: OpenData RATP 2014) in the Île-de- France region, RATP develops its activity and exploits its infrastructure in a highly dense area. It is interesting to compare the rules for building national timetables described previously with those of an urban and outer-urban transport manager operating in a very dense perimeter. The timetable construction rules described in this section correspond to the RER network. A comparison with the construction of the subway timetable does not appear pertinent since it is very different in terms of technology, rolling stock and utilisation.

Service design: timetable design and the margins applied

The details of the construction standards of the graphic timetable at RATP were provided from exchanges with Mr Patrick Bonan, RER project manager.

The design of the RER supply was initially performed by reproducing the method used to design the subway supply. Different constructions of supply exist for the type of day (weekday, weekend, holidays, etc.) and a total of more than twenty standard days are available.

The RER travel times designed by RATP are all counted with a margin of 4% in addition to the theoretically possible travel time using the least efficient rolling stock of the line. This is done to take into account possible dispersion in driving behaviour (mainly by anticipated braking and irregular observance of speed limits).

During peak hours, the RER lines are used to their maximum capacity with a headway between trains corresponding to the minimum technical headway. Theoretically, timetable construction takes into account the visibility of the "free block" for drivers. To do this RATP builds the graph with a 15-second margin, that is to say that the signal must have changed to "free block" 15 seconds before the train crosses it so it can be seen by the driver. The value of this distance contrasts with the national network standards for layouts that provide for a "khi" of 35 seconds. In normal operation, the RATP's "khi" is not complied with today on the shared line A/L3. The reality of the number of trains required in this area sometimes makes it necessary to override these additional headway rules, thereby ensuring robustness.

In the case of the RER (lines A and B), which are the two most heavily used railway lines in Europe, with 1.2 million and 870,000 passengers, respectively per weekday, the trade-off between capacity and robustness of the system is clearly made in favour of capacity, given the demand and the strong need for mobility. During peak hours, the supply programmed for RER line A is 30 trains an hour (thus a headway of 2 min between trains) and 20 trains an hour for RER line B (thus a headway of 3 min between trains).

Nonetheless, the figures on service quality published quarterly by STIF reflect the cost in terms of the punctuality of this intensive utilisation of capacity. For example, the objective for regularity set by the contract between STIF and RATP is 94% [1] although

[1]Passenger punctuality represents the percentage of passengers arriving on time or 5 minutes late at their destination station. Passenger punctuality is calculated for the whole line and throughout the day. It is based on the timetable displayed in the stations and on theoretical passenger flows.

the figures of the second quarter of 2015 show punctuality rates of 85% and 89 % for RER lines A and B respectively.

RATP does not test robustness systematically. In comparison to the robustness standard applied for the national network and described in the Statement Document of SNCF-Réseau[2], RATP states that when a delay of 10 minutes cannot be absorbed, it leads to a 30-minute delay one hour later if no operating measure is taken. Daily operation is considered as a test upon which timetable building can be adjusted empirically, by trial and error.

1.5.2 Prorail

ProRail is the public body responsible for managing the national railway infrastructure of the Netherlands. Its missions are maintaining the lines and installations, allocating capacities and traffic management.

The rules relating to the timetable construction process are described in their Network statement (2017). Generally, its planning consists in calculating a minimum technical time as a function of the infrastructure and the characteristics of the rolling stock and by considering additional times. This practice is similar to those described previously for other networks, but in this case all the additional times are described in a publicly accessible document (Network Statement Prorail, 2017).

Service design: timetable design and the margins applied

The details of the construction rules of the graphic timetable at ProRail results from exchanges with Mr Vincent Weedaand Jan Swier, a rail traffic analyst at Prorail.

To determine travel times, Prorail first uses the "Donna'" software that calculates the minimum technical times between two blocks (including the durations of stops if necessary). To this basic running movement, the Dutch IM adds 5% regularity margin for all the passenger trains and the planned travel time is rounded off to the highest figure. Regarding freight trains, basic running movement corresponds to the planned travel time (the regularity margin is therefore equal to 0%).

In addition, Prorail calculates the headway and crossing times between two trains (for both passenger and freight trains) and rounds them off to the nearest minute. An additional minute of headway is systematically added to this minimum time interval

[2]It should be remembered that a delay of ten minutes on a train should be absorbed after one hour

for all types of traffic. According to exchanges with the persons responsible for timetable construction in the Netherlands, although the general rule is to have an additional minute of headway between two trains, in practice it can vary between rounding off at 0.5 or 1.5 minutes. By way of example, a technical headway slightly longer than 2 minutes will result in a planned headway of 3 minutes, thus an additional headway of less than 1 minute.

The margins applied today result from a process that includes analyses and learning from critical situations. The choices made by ProRail relative to regularity margins and robustness are based on feedback from the field. According to ProRail, the trade-off between capacity and punctuality currently applied appears reasonable, although it has not been subjected to economic formalisation.

1.5.3 Infrabel

Infrabel is the Belgian infrastructure manager. It is in charge of the maintenance and renewal of railway infrastructure as well as extending the capacity of the railway infrastructure as a function of mobility requirements. It organises the operation of the railway infrastructure and the distribution of available capacity between the railway companies, and the daily coordination of all the trains running on the Belgian railway network.

The details of the standards used for building the graphic timetable at Infrabel result from exchanges with Mr Axel de Bie Gaona, Long-Term Timetabling analyst.

The underlying timetable design is very similar to that described for the previous networks. Nonetheless, the values of the margins applied are different from those of the networks described above.

The regularity margin applied individually to trains is 5% for passenger trains and empty trains, and 7% for freight trains. Infrabel adds (excluding HSL) 1 min / 35km to these margins (at the discretion of the timetable manager).

Regarding headways between trains, Infrabel generally applies a minimum headway of three minutes between trains, and avoids placing more than four successive trains with a minimum headway. This practice is similar to that of the “buffer train path” described in the case of SNCF Réseau.

Passenger timetables are adapted on the basis of a transport plan, valid for at least 3 years. This timetable may be adapted seasonally (and daily). Before a transport plan

is implemented, it is analysed from the standpoint of robustness using a simulation tool (LUK-S).

Seasonal timetables are also analysed during production. Records of real traffic are regularly studied and variances are subjected to proposals for timetable improvement or modification.

1.5.4 Ferrocarrils de la Generalitat de Catalunya(FCG)

The railway network of the Government of Catalonia has a number of lines both in urban and suburban areas, supplying specific intervals for metropolitan services in the city of Barcelona, and semi-direct trains for cities outside Barcelona. Timetable design is based on programming service supply to match the demand of passengers at different time periods and is compatible with the network signalling and protection systems.

The principles of the timetable construction process at FGC

FCG has established the following criteria for itinerary timetable design:

- Maximum supply at peak hours in relation to the theoretical capacity of the line and the available rolling stock.
- Regularity margins and additional times when designing train movements and the rotation of rolling stock.
- Clock-face timetable.
- Synchronisation of departures from origin stations to optimise hub management and minimise connection times.
- Balance between the supply of urban and suburban sectors according to demand during time periods.
- Optimisation of travel time by conforming to regularity margins and additional times.

Once the timetable has been built and before it is used, FGC performs a dynamic simulation of movements to check the robustness of the service planned.

Service design: timetable design and the margins applied

The details of the standards applied for graphic timetable construction at FGC result from exchanges with Mr Oriol Juncadella i Fortuny, director of FGC Operator.

In order to clearly understand the timetable construction process, it is necessary to describe the concepts used in calculating the layout by FGC:

Maximum speed: this is determined by the layout (geometry, infrastructure, signalling) of the journey and the rolling stock. This speed cannot in any way be exceeded by the train.

Route speed: this corresponds to the speed of trains between different stopping points in phase with the "allotted time", to the exclusion of starting and stopping times at stations.

Allotted time: this corresponds to the time calculated between two stopping points when travelling at the speed of the route. This time includes the regularity margin.

Regularity margin: this is defined as the time difference between circulation at maximum speed and circulation at the speed of the route for the same trip.

The running movement of trains is calculated using a route speed less than the maximum speed. Initially, the minimum travel time between two stopping points is calculated theoretically by considering the maximum speed. To this theoretical travel, FGC adds an additional time close to 50 seconds/10 km to obtain the time allotted for a given journey and thus the route speed.

Additional margin: this corresponds to an additional time added to the minimum station stopping time in certain circumstances. FGC considers stops to last 20 seconds. These times can be increased in certain cases:

- In stations with high volumes of passengers, for example the station of Provença (station connecting with two TMB subway lines), this stopping time is set at 55 seconds.

- In stations with crossing tracks, junction branching forks or the convergence of different lines.

- In terminal stations, an additional time can be added to the minimum train turnaround time. This time is designed to ensure the stability of the network by minimising the impact of a delay of a mission on the following missions. By way

of example, this additional time is situated on the FGC network (Barcelona-Valles line) at between 5% and 7% of the route time of one train movement on a line. This additional margin mainly stems from experience in the field on the real punctuality of trains, and is influenced by the availability of tracks and rolling stock (an increase in turnaround time can involve an additional need for rolling stock for a given transport plan).

Analysing the stability of a journey

At the end of the service design process, FGC carries out an evaluation of timetable stability/robustness using the OpenTrack software®(EPFL). This software simulates the behaviour of a railway service on the basis of an infrastructure, rolling stock and times fixed previously.

To determine whether a service timetable is stable, FGC carries out robustness tests based on Pachl's (2009) analyses. This method consists in generating a delay of 10 minutes in the most difficult section of the line and checking that the system evolves according to following conditions:

1. The sum of delays recorded at the exit of the system (the trains that exit the system and the trains that end inside the system) is lower than the sum of delays introduced in the system (the trains entering the system and the trains that start in the system).
2. The theoretical timetable of the service is restored after two full cycles at the latest[3] for each circulation.

The second condition is calculated on the basis of the following formula:

$$q_{resilience,j} = \frac{t^i_{j,k}}{t^{reg}_{j,k} + t^{rec}_{j,k}} \tag{1.1}$$

where: $t^i_{j,k}$ is the delay of a train that enters the section studied, for a given movement j, and a cycle k of the movement considered, $t^{reg}_{j,k}$ is the margin of regularity of movement j for cycle k and $t^{rec}_{j,k}$ is the additional margin of movement $j$for the associated cycle k.

In order to check that the delay of 10 minutes is completely absorbed at the latest after two full cycles for all the movements of the period analysed, and thus satisfies the minimum condition:

[3]A full cycle corresponds to the total time between two successive departures between the same set of trains

$$q_{resilience,j} \leq 2 \tag{1.2}$$

The quotient $q_{resilience}$ considers all the delays of the trains from the moment the incident is generated.

1.6 Conclusion

This first chapter on the characterisation of the notion of railway capacity from an engineering point of view, shows that reflection abounds in this area, from both the academic and industrial standpoints.

Prior programming of railway supply determines the nature of the capacity constraints affecting railway transport. The definitions of railway capacity that can be found in the academic literature facilitate understanding the complexity of its nature on the one hand, and the importance of having adapted measurement tools on the other.

One of the key parameters for defining railway capacity is the level of service quality required. Analysis of the different timetable production processes employed in France and elsewhere in Europe, shows that railway infrastructure managers are aware of and include the link between the capacity and the robustness of their operations in their timetable choices. Nonetheless, differences can be seen in the practices employed by infrastructure managers, by country and by type of network. Generally, the standards of robustness applied for the different networks are often tacit and based on empirical feedback. Infrastructure managers appear to proceed by trial and error to define certain of their timetable design rules, which leads to practices designed empirically.

Nonetheless, existing analyses do not allow considering the value of capacity constraints in the rail infrastructure manager's decision-making process. In this context, it seems relevant to start reflection on railway capacity constraints from the economic standpoint, in view to objectifying the previous trade-off.

Chapter 2

Theoretical Economic Framework

Contents

2.1 Introduction

Once the concept and the elements of capacity have been analysed from the angle of engineering, our aim in this literature review chapter is to develop the theoretical framework in which we can study the economic design of the railway capacity constraint.

Many networks suffer from peak-load demand problems, meaning that individual user behaviour has an impact on the costs of other users, creating an externality. In general, congestion refers to the existence of limited capacity networks for which demand varies periodically and their intensity of use impacts the quality of service. As stated in the introduction there is a large amount of engineering and empirical literature that relates to the modelling approaches implemented and the types of congestion technology, mainly intended for the road sector. This chapter focuses on the economic approach of capacity constraints.

Economic analysis has oriented research towards studying the link between the quality of service and the degree of utilisation, as previously stated in the engineering literature. Once empirically verified, economists attempt to understand and formalize the consumer's behaviour that leads to congestion and its consequences for other users.

Since Pigou (1920) used the example of a congested road to explain the economic concept of external effects, a considerable amount of literature has worked on road congestion, like the major contributions from Knight (1924), Wardrop (1952) or Walters (1961).

2.2 Static Models or Classical Contributions

Analysis using static models of traffic congestion is mostly used for research or educational purposes; static models are a basic tool for the mathematical description of congested networks. In classical contributions on congestion, time is not explicitly considered, which means that they might overlook changes in congestion over time, like during peak and off-peak periods.

2.2.1 Short-term models

Following the description by Lindsey and Verhoef (2000), the basic principles of road congestion and the corrective "pigouvian" tax can be illustrated in the following description. Consider a single road connecting a pair of cities A and B. The users are identical (same travel time costs, same vehicles) and they travel alone. A particular characteristic of

static models is that traffic flow, speeds and densities are time-independent and uniform along the road.

At low levels of traffic, vehicles can travel at a free-flow speed for a constant average variable cost $C(q)$, but when traffic increases and reaches maximum basic capacity, $C(q)$ slopes upward due to significant negative interactions between users, increasing congestion [1].

Once the cost curve is set and in order to establish a supply-demand diagram, we consider an inverse demand function $p(q)$. The inverse demand function is assumed to slope downwards to reflect that the quantity of trips demanded decreases with cost. The inverse demand function reflects users' willingness to pay and their marginal benefits of travelling. At equilibrium, users equalize their willingness to pay $p(q)$ with their generalised cost of travelling gc, which is defined as the average variable cost $C(q)$ plus a possible toll τ.

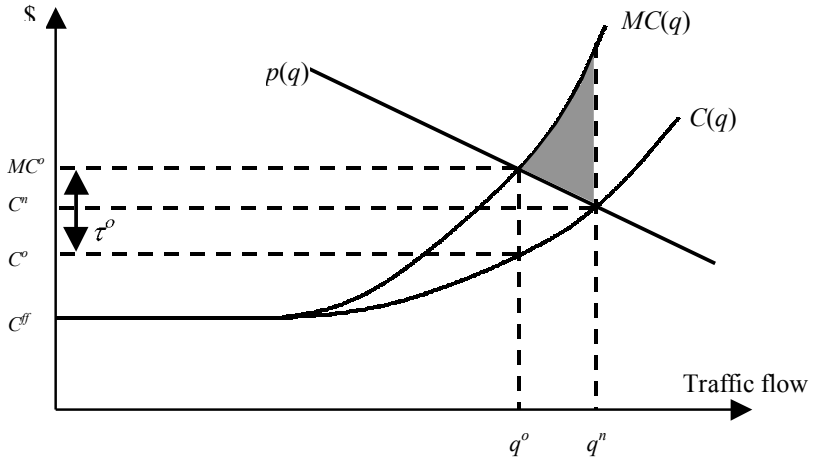

Figure 2.1: Optimal road pricing in a time-independent model. Source: Lindsey and Verhoef (2000)

In figure 2.1, the horizontal axis illustrates the traffic flow and the vertical axis depicts the generalised cost for a trip, considering vehicle and time costs $C(q)$ and any toll τ.

This figure also shows a private equilibrium point at flow q^n and price C^n, where users' marginal benefit of travelling equals the average variable cost curve (the function upon which travellers base their trip decisions). As Hau (1998) recalls, basic price theory establishes that whenever the average variable cost rises $C(q)$, the marginal cost curve

[1] This description ignores the possibility of "hypercongestion"

$MC(q)$ must lie above it. Formally, the social total cost of q trips considers the average variable costs $C(q)$ and thus $TC(q) = C(q)q$. Therefore the marginal social cost of an additional trip is $MC(q) = \partial TC(q)/\partial q = C(q) + q\partial C(q)/\partial q$.

The vertical difference between the average cost curve $C(q)$ and the $MC(q)$ curve is the marginal external congestion cost, namely the additional delay that one user imposes on the other drivers. Indeed, individuals take travel decisions by considering their private cost (which corresponds to $C(q)$), but they completely disregard the additional cost that they impose on other drivers.

On the other hand, the social equilibrium obtained at the intersection of $MC(q)$ and $p(q)$, associated with an optimal output q^o and price MC^o, takes into account the external congestion cost and other variable costs. We note that the first output equilibrium q^n is higher than the social equilibrium q^o which considers all costs.

To obtain q^o as the number of trips at equilibrium, users should pay the total price of MC^o. Optimal charging should consider the additional time that one driver imposes on others: $\tau^o = MC(q^o) - C(q^o) = q^o q\partial C(q^o)/\partial q^o$. The Pigouvian tax introduced by Pigou (1920) and applied to roads is the toll that erases the gap between the marginal cost and the average variable cost curves by issuing the correct signal and creating appropriate incentives.

As stated by Small and Verhoef (2007) the social optimum q^o does not mean the absence of congestion (the generalised cost net of the toll, C^o is higher than the free-flow cost C^{ff}). At equilibrium, some congestion is also considered as optimal. The gain in social surplus considering the toll equilibrium is given by the shaded "Harberger" triangle, which is defined by the difference between the reduction of social costs (the area below $MC(q)$) and the reduction in total benefits due to the decrease in traffic (the area below $p(q)$).

To sum up, when car users decide to make an additional trip, they impose additional costs on themselves, on the infrastructure provider and on other users. From the economic perspective, congestion is basically a standard externality problem. Academic literature shows that peak/off-peak pricing is an efficient solution for tackling congestion and obliges users to internalize the external costs generated.

In practice, there are very few examples of cities that have implemented a congestion toll, possibly related to the lack of public and political acceptability and to the highly political discussion on distributional equity. The most well-known examples are

the High Occupancy Toll lanes in the United States, the Singaporean Electronic Road Pricing system, the London Congestion Charging Scheme and the Stockholm congestion tax (Santos, 2004). As Santos et al. (2011) describes, most of the schemes implemented have achieved the objectives targeted: decreasing traffic in some areas, improving travel time and collecting net revenues designated for new road investments. Nevertheless, they also remark that none of the congestion pricing schemes applied was designed according to economic rules (first-best or second-best). The authors concluded that *" the schemes in operation are therefore not so much a triumph of economics as of political will, or at most, of political determination somehow inspired by economic ideas".*

2.2.2 Long-run models

In the static models, the short-term approach investigates the economics of congestion with fixed capacity and optimal pricing. In order to complete the analysis of congestion, it is now necessary to consider capital investments as an additional adjustment variable in long-run congestion management.

Capacity choices for infrastructure are an essential step in congestion analysis, combining optimal pricing and optimal investment in the same methodological framework.

As in the short-run analysis, significant lessons can be drawn from the basic static congestion model. In the previous section, welfare maximisation in the static model depended on total social benefits and total social costs, but omitted investment capacity cost. Following the formulation of Mohring and Harwitz (1962), $K(S)$ characterizes the relationship between infrastructure size S and capital investment expenditures.

In the long-run, social welfare can be written as:

$$W = \int_0^D F(q)dq - Dg(D,S) - rK(S) \tag{2.1}$$

where D is the vehicle flow, S road size and r the optimal interest rate for public investments.

The first order conditions for maximising long-run social welfare can be found by maximising W with respect to D and S.

The first differentiation $\frac{\partial W}{\partial D}$ gives the marginal-cost pricing rule presented previously. In order to obtain the optimal investment rule, we maximize W with respect to road size

S, which leads to:

$$-D\frac{\partial g}{\partial S} = r\frac{\partial K}{\partial S} \tag{2.2}$$

At the optimum, the marginal capital cost of incrementing road size (thus capacity) is equal to the saving in marginal congestion cost, provided by an increment in size.

Mohring and Harwitz (1962) ask *"the question of immediate relevance is, under what conditions will the optimum capital charge* (rK^*) *equal optimum annual toll collections* $(D^*g_D^*)$*?"*. We can in other words wonder, under which circumstances will congestion fees generate enough revenue to cover the cost of incrementing capacity, and by consequence be self-financed.

As demonstrated by Mohring and Harwitz (1962) and summarised by Small and Verhoef (2007) the self-financing result applies when certain restrictive conditions are fulfilled: (a) constant returns to scale in congestion technology (doubling traffic flow and road size would mean the same congestion costs for drivers), (b) neutral scale economies in capacity provision (the cost for providing a road with four-lanes and a two-way double lane road is exactly the same), (c) perfect divisibility of capacity (a condition that is not explicitly named in the seminal paper of Mohring and Harwitz (1962) but which is an implicit assumed condition).

The initial analysis of self-financing results has been extended in different directions. The objective of these extensions has been to consider a number of initial assumptions and verify whether the seminal self-financing rule remains valid and with which deviations.

Our description concerning the extensions follows previous reviews such as those of Hau (1998), with an extensive diagrammatic analysis, De Palma and Lindsey (2007) and Small and Verhoef (2007).

First, the perfect divisibility condition assumed by Mohring and Harwitz (1962) may seem unrealistic. Road construction implies significant indivisibilities that must be considered. Figure 2.2 illustrates the problem more generally. When indivisibilities exist, the short-run average cost (atc) follows a U-shape. When the short-run marginal cost (srmc) is lower than the average cost (downward sloping segment of the atc), the operator will generate a deficit if it applies a pricing rule equal to its marginal cost. On the other hand, the result will be a surplus if the short-run marginal cost is higher than the average cost (upward sloping segment). Without indivisibilities, the long run marginal cost (lrmc) curve would be a horizontal line, and the long-run average total cost (lratc) curve would be downward sloping.

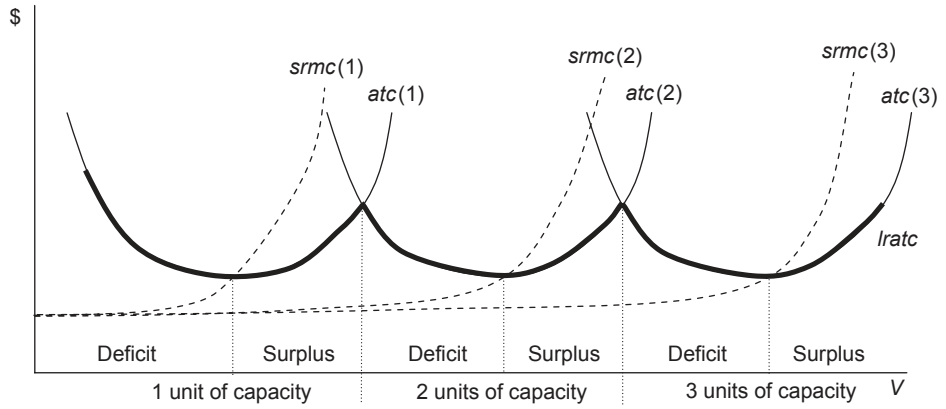

Figure 2.2: Surplus and deficits. Source: Small and Verhoef (2007)

The logical issue which follows is to determine if indivisibilities cast doubt on the self-financing theorem. As both Verhoef and Mohring (2009) and Small and Verhoef (2007) observed, this depends on the situation. In areas with low demand and no congestion issues, the (no) congestion revenues may not allow financing capacity investments. Nevertheless, if demand grows over time, and capacity is periodically increased, surplus and deficit periods will alternate and offset each other (or the discounted net deficit or surplus will be small). Also, if a network with many roads is considered, some roads will generate surpluses and other deficits, and possibly cancel each other when aggregated.

2.3 Dynamic Models

Static models are considered very useful for explaining and understanding the basic economic mechanism of traffic congestion and the costs and benefits of a corrective toll. Nevertheless, static models omit some characteristics usually observed in real traffic congestion diagnoses, such as the fact that congestion varies over the day, with peak-period demand in metropolitan areas.

As mentioned by De Palma and Fosgerau (2010), the characteristics of demand peaks should be taken into account by congestion models. Congestion economics models are based on knowledge of traffic engineering research.

First, De Palma and Fosgerau (2010) considered that the departure time choice of users is an important variable when congestion varies over the day. Users are able to modify and adjust their departure time if congestion policies are implemented. Secondly,

they recall that travellers have preferences regarding the time of their trips. Static models omit scheduling costs, considering that delay cost and travel cost are the only travel time costs. In fact, the authors consider that dynamic models alone are capable of describing congestion policies and revealing their real impact on the user's total costs.

In transport research, morning and evening peak-hour congestion is considered as a classic problem of trip scheduling under deterministic traffic conditions. Vickrey (1969) presented the first model with a single deterministic bottleneck which was further extended by Arnott et al. (1990), Arnott et al. (1993), Chu (1995) or Verhoef (2001).

The "basic Vickrey bottleneck model" considers an inelastic demand $N > 0$ [2] of identical travellers who have to pass through a bottleneck with a constrained capacity s. As shown in Figure 2.3, the bottleneck is located d_1 units from the trip origin and d_2 time units from the destination. Users arrive at the bottleneck at time t and exit at time a. If traffic inflow is below capacity, there is no delay and all travellers can pass through the bottleneck. If not, travellers will spend some time in the bottleneck and will suffer delays.

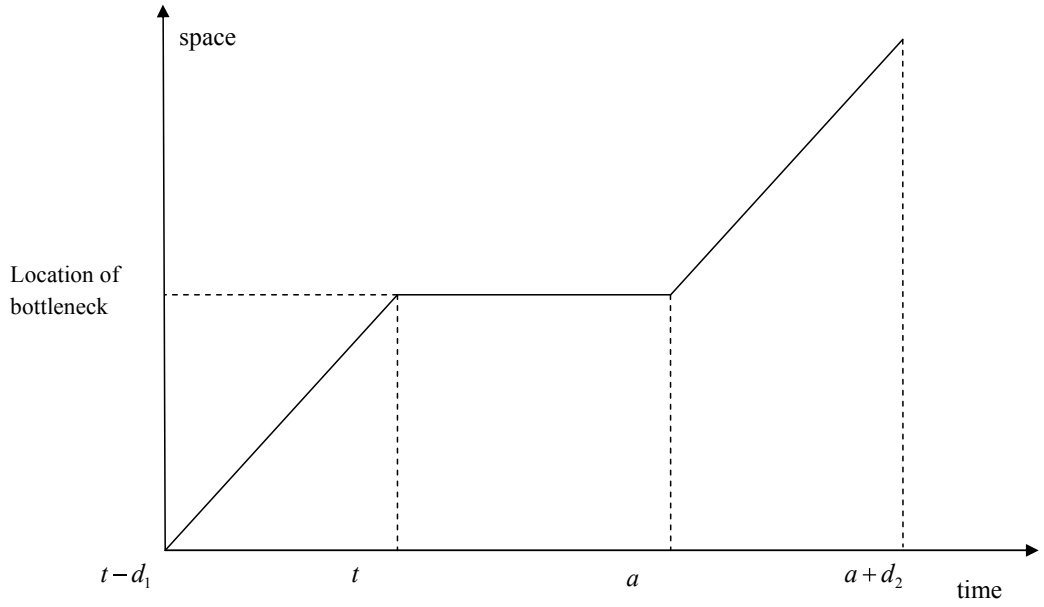

Figure 2.3: Trip timing. Source: De Palma and Fosgerau (2010)

The basic model considers that each user has preferences concerning the timing of their trip and their aversion to arriving earlier or later. Their preferred arrival time is t^*.

[2] In this description we follow the notation by De Palma and Fosgerau (2010)

Travellers do not like spending time in bottlenecks and extending their travel time. When a user decides on a departure time t_1, they take into account scheduling and delay costs (depending on the bottleneck). For a trip that starts at time t_1 and ends at time t_2, the user travel cost is:

$$c((t_1, t_2) = \alpha(t_2 - t_1) + \beta max(t^* - t_2, 0) + \gamma max(t_2 - t^*, 0) \tag{2.3}$$

The cost parameters are assumed to be negative and identical for all users. In this formulation, α is the value of travel time, β and γ are the shadow prices of early and late arrivals compared to the preferred arrival time t^*.

The travel time between the origin and the bottleneck and the bottleneck and the destination is usually set to zero, meaning $d_1 = d_2 = 0$.

R is the cumulative departure rate, where $R(a)$ is the number of travellers departed before time a. The bottleneck can serve a maximum of s travellers per unit time. If the departure rate is higher than the bottleneck capacity, some travellers must queue before the bottleneck.

As users are considered to be identical, all travellers will be subject to the same total costs. Nash equilibrium conditions (defined as a situation in which no traveller is able to decrease his cost by choosing a different departure time) for the departure and arrival interval $I = [a_0, a_1]$ are defined by:

$$a_1 - a_0 = N/s \tag{2.4}$$

$$\beta(t^* - a_0) = \gamma(a_1 - t^*) \tag{2.5}$$

Equation 2.4 shows that the interval depends on the time needed to pass through the bottleneck (as a function of its capacity and the number of users). Equation 2.5 demonstrates that travellers are not interested in departing at another time outside I.

Solving these equations gives the peak start and end times:

$$a_0 = t^* - \frac{\gamma}{\beta + \gamma}\frac{N}{s} \tag{2.6}$$

$$a_1 = t^* + \frac{\beta}{\beta + \gamma}\frac{N}{s} \tag{2.7}$$

at equilibrium, the cost for every traveller is:

$$\frac{\beta\gamma}{\beta+\gamma}\frac{N}{s} \equiv \delta\frac{N}{s} \tag{2.8}$$

Consequently the total cost is $\delta\frac{N^2}{s}$ and the corresponding marginal cost following a change in users is $2\delta\frac{N}{s}$. If there is no toll, the generalised price equals the private travel cost $\delta\frac{N}{s}$.

A dynamic equilibrium is defined as the situation in which no user can reduce his or her costs, by unilaterally changing the departure time from home. As travellers are all identical, they incur the same scheduling cost in equilibrium during the interval I:

$$\delta\frac{N}{s} = \alpha\frac{R(a)}{s} + \beta max\left(-a_0 - \frac{R(a)}{s}, 0\right) + \gamma max\left(a_0 + \frac{R(a)}{s}, 0\right) \tag{2.9}$$

being $\frac{R(a)}{s}$ the total time to pass through the bottleneck for all the travellers $R(a)$. Differentiating this expression makes it possible to obtain the departure rate during the interval:

$$\rho(a) = \begin{cases} s\dfrac{\alpha}{\alpha-\beta}a_0 + \dfrac{R(a)}{s} \leq 0 & (2.10) \\ s\dfrac{\alpha}{\alpha+\gamma}a_0 + \dfrac{R(a)}{s} > 0 & (2.11) \end{cases}$$

Graphically, as can be seen with the departure rate at the beginning, the number of departures is higher than the capacity and a queue builds up. The queue length corresponds to the segment $b - c$, i.e. the difference between the cumulative departures and the numbers of travellers served by the bottleneck. The first users are subject to an increasing queue cost and they arrive earlier than the preferred arrival time t^*. The user departing at time d will arrive exactly at their preferred arrival time t^* but will be subject to the maximum length in the bottleneck queue. After d, the queue starts to diminish. Later travellers (after d) will spend less time in the queue, but will arrive later at their destination.

The equilibrium of non-coordinated travellers' decisions generates a travel queue delay cost that is a pure dead weight loss: nobody benefits at all. If it were possible to coordinate or induce travellers to depart at the capacity rate s ($\rho(a) = s$), no queue would form, but they would arrive at the destination at the same time as they did in the previous equilibrium. The principal lesson of the bottleneck model is that a new optimal

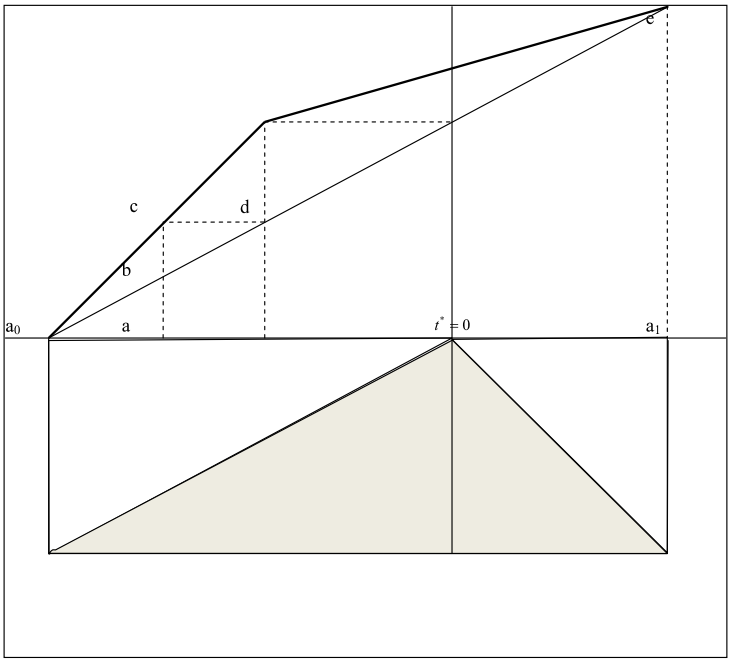

Figure 2.4: Equilibrium departure schedule. Source: De Palma and Fosgerau (2010)

equilibrium without a queue can be achieved by using a toll. The toll pattern must be exactly the same as the travel delay cost at the queue in the no-toll equilibrium.

The optimal toll is:

$$\tau(a) = \delta\frac{N}{s} - \beta max(-a, 0) - \gamma(a, 0) \tag{2.12}$$

Indeed, travellers do not gain or lose with the implementation of a toll (their generalised cost is the same, replacing the queueing cost by a toll, and they arrive at the same time as they did before). Nevertheless, implementing a toll generates income for the road owner that could be used for other purposes.

The seminal paper of Vickrey (1969) has been considered in a large number of articles and the model it proposes has been applied with a number of extensions.

Cohen (1987), Newell (1987) and Arnott et al. (1988), extended the initial model allowing heterogeneity for drivers, with different preferred times t^* or different values for the scheduling parameters.

Arnott et al. (1990) formalised the Vickrey model and extended it to consider a coarse toll and solve optimal capacity. Indeed, implementing the fine toll described above, which exactly replaces the queueing cost for each user, implies exhaustive knowledge of the

parameters δ, γ, β,N/s and t^*. Consequently, the authors proposed a coarse toll, i.e. a toll with a single step during peak hours. The aim of this extension was to determine the optimal toll and the time interval for its implementation. The paper demonstrated numerically that a significant proportion of the gains associated with a fine toll can also be achieved by a coarse toll, which is simpler and less costly than the former.

In this article, Arnott et al. (1990) also detailed the desirable extensions capable of leading to a full economic description of peak hour phenomena and which have been the basis of further developments.

The demonstration of the initial bottleneck model assumes that demand is inelastic. Arnott et al. (1993) extended it to consider the case with elastic demand, where drivers' decision to travel depends on their trip cost and the optimal capacity choice under different toll regimes.

Besides the bottleneck model, alternative sophisticated dynamic congestion functions also exist such as that of Chu (1995). Small and Verhoef (2007) compared alternative functions with the basic bottleneck model and concluded that "*the bottleneck model overestimates the benefits from optimal tolling, and underestimates the resulting increase in generalised price, by exaggerating the extent to which travel delays can be eliminated without increasing scheduling costs*".

Until now, the optimal tolls and the conclusions concerning marginal pricing in the previous sections have been referred to as "first-best" solutions. In fact, these recommendations do not consider additional market distortions and practical constraints in their implementation. First-best analysis provides important lessons concerning congestion prices but, in reality, market distortions exist that call for second-best analyses. In practice, second-best analysis allows describing and analysing the best policy options in a constrained world. For a further discussion on second-best analyses an extensive review can be found in Small and Verhoef (2007).

As in static models, a natural extension of the dynamic short-term optimum is to consider long-term congestion models or user heterogeneity. Arnott and Kraus (1998) considered these two aspects and showed that the general results of the self-financing theorem remains valid under certain conditions.

Indeed, economic research on dynamics models follows the same pattern as that of static models. In general there is a seminal model with simplified hypotheses that allows illustrating significant concepts from a theoretical point of view. Furthermore, this initial

model permits extensions and becomes more refined when certain assumptions are relaxed. Finally, this theoretical research integrates additional market distortions and practical constraints in its reasoning in order to provide realistic policy recommendations (second-best world).

Dynamic equilibrium congestion models based on the concept of "scheduling preferences" are particularly interesting because they explain passenger behaviour when travelling and the associated costs.

Since individual behaviour is a major input in dynamic equilibrium models, a complementary path of research based on detailed analysis on individual demand has been developed over the last thirty years. These works describe the individual behaviour characteristics (under different conditions) that underlie congestion technology and their main objective is to explain the departure time choice of users and its costs under different conditions.

2.4 Individual Behavioural Models

In order to plan efficient transport services it is important to understand how much transport services are going to be used and under which conditions.

In view to better planning transport policies it is necessary to know passengers' behaviour characteristics and how they react to changes in prices and service quality aspects.

Standard micro economic analysis describes and analyses demand and supply functions and seeks an optimal equilibrium between them. The particularity of transport microeconomic analysis is the role given to the passenger. On the one hand, passengers are consumers of transport services (as in standard consumption markets), but on the other hand they are also producers, as their time is an input of the transport production function.

Consequently, the transport demand function must consider all the costs incurred by passengers when travelling: monetary cost (fares, vehicle maintenance costs, tolls and parking charge) as well as non-monetary costs such as time spent travelling. In transport economics, the generalised cost is the sum of the monetary and non-monetary costs of a journey.

Travel demand forecasting can be done using aggregate or disaggregate models. In the first case, the variable studied is the total demand for a particular market, considering all

the variables and characteristics that describe the product (income, quality characteristics, costs, services, travel time, etc.). In the second case, disaggregate demand models explain individual behaviour using micro-data (based on individual decisions). Most of these estimations are based on discrete-choice models whose theoretical foundations mainly stem from McFadden (1974) .

Travel time can be defined as the time spent when a traveller moves between two different places. Moreover, travel time can be split up into different components depending on the objective of the analysis. For example, travel time in public transport is usually divided into waiting-time, in-vehicle time and transfer time. In road networks, travel time can be split into two components: free flow time and additional time (Carrion and Levinson (2012)).

In transport where timetables are planned in advance (like rail transport), travel time can be broken down into three components: planned travel time, expected schedule delay cost (individuals travel either earlier or later than they would like to), and a random delay cost if the vehicle arrives later than expected in the timetable.

The value of travel time (VTT) is one of the cornerstones of transport economics research. The VTT concept allows analysing travel behaviour and it is an essential variable in traffic assignment models. It is also an important element in CBA analysis, where VTT savings are the main benefit derived from transport investments.

Becker (1965) probably wrote the seminal paper which explained consumer behaviour, by considering the allocation of time for multiple activities and considering its value. Since then, the concept of VTT has been introduced in the utility functions of different activities including the transport sector via travel time cost. How individuals decide to carry out activities is an important feature for understanding travel demand distribution during the day.

An important contribution to the development of the travel utility function was the introduction of activity scheduling in the analysis. Departure time choice is an important element in travellers' decision-making. Usually, it takes into account consumers' preferences: waking up later, having breakfast at home, arriving first at office, etc.

In road transport, travellers are free to choose their departure time. On the contrary, in public transport, travellers can only choose between fixed scheduled services, as defined by the timetable. Travellers' departure choice will influence their arrival time, assuming that there is a disutility in early or late arrivals and thus in their travel cost (or utility function).

Economic formalisation of individual travel decisions can also be classified into two categories: under deterministic or under stochastic conditions.

2.4.1 Individual behaviour under deterministic conditions

Determining how to analyse and explain a traveller's departure time choice behaviour and the associated travel cost has been one of the main paths of research in transport economics.

The initial research carried out by Vickrey (1969) presented a single bottleneck model illustrating that peak congestion is a classic problem of trip scheduling choice under deterministic traffic conditions. This paper allowed understanding the user trade-off between the queue delay and the schedule delay of arriving early or late at work before choosing an optimal departure time under deterministic assumptions.

In deterministic approach models, consumers are assumed to be fully informed and there are no unreliability problems. Commuters need to arrive at work before the start-time in order to avoid a penalty, but their standard travel time is associated with travel time variability.

The aim of the trip scheduling model is to understand the choice of departure time when travellers face time constraints associated with work-start time. As pointed out by Li et al. (2010), the scheduling model considers that disutility is incurred when one does not arrive at the preferred arrival time (PAT), either early or late.

Based on the earliest research on this concept performed by Gaver Jr (1968) and Vickrey (1969), another essential contribution to this framework was that of Small (1982). He explicitly estimated the utility function parameters and detailed a preliminary theoretical linear model which has been extensively used in theoretical works:

$$U = \alpha E(T) + \beta E(SDE) + \gamma E(SDL) + \theta D_L \tag{2.13}$$

The official work-start time determines the trip scheduling decision. In Small's model formulation, T is considered as the travel time and the schedule delay S_D is defined as the difference between the arrival time and the official work-start time. The schedule delay can be broken down into two terms: Schedule delay early, SDE as Max {-S_D, 0} or Schedule delay late, SDL as Max {S_D,0{. D_L is a dummy variable equal to 1 when there is an SDL and 0 otherwise. The estimated parameters (α, β, γ and θ) correspond to the

time-related shadow prices of travel time, arriving early and arriving late, respectively. They are assumed to be negative.

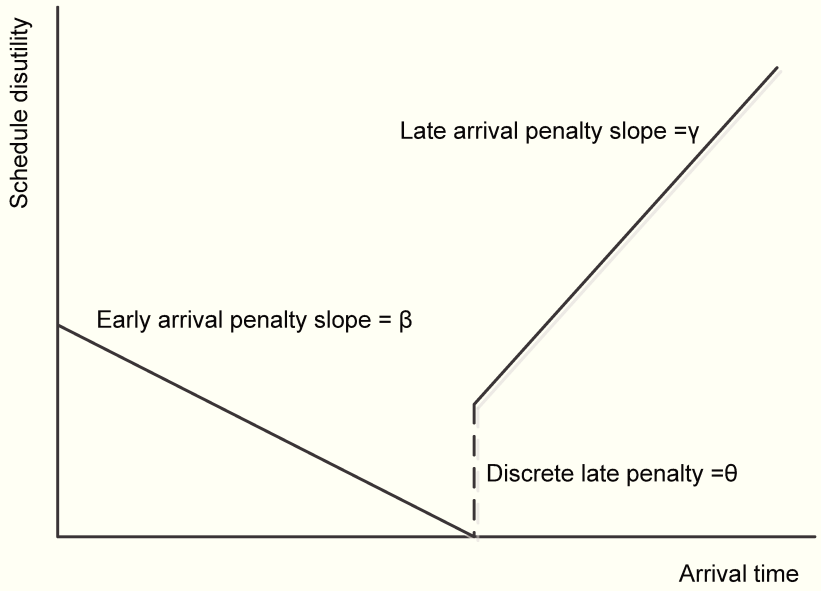

Figure 2.5: Small's Formulation of Arrival Delay Disutility. Source: A. Nour (2009)

As described above, the formulation of Small (1982) assumes a linear utility specification. Tseng and Verhoef (2008) and Fosgerau and Engelson (2011) proposed a variant of Small's model, by generalising the scheduling preferences model with a non-linear form.

2.4.2 Individual behaviour under stochastic conditions

In order to consider travel time reliability, the initial scheduling model has been further developed considering stochastic travel time , e.g.,(Noland and Small, 1995; Bates et al., 2001; Fosgerau and Karlström, 2010; Fosgerau and Engelson, 2011).

Currently, we can distinguish three methodological approaches in the literature to estimate and measure travel time and its variability (Li et al., 2010):

- Mean variance approach: travel time variations are considered as a cost in the consumer utility function. Travel time variability can be represented by the variance or the standard deviation of travel time. Thus the formulation (with a linear additive form) of the model, with a consumer theory background, is as follows:

$$E(U) = \beta_T E(T) + \beta_{SD} SD(T) + \beta_c C \tag{2.14}$$

where β_T, β_{SD} and β_c are the estimated parameters for the expected travel time $E(t)$, the standard deviation of travel time $SD(T)$ and the travel cost C respectively.

Following Benezech and Coulombel (2013), we can consider that the "Mean Variance" is a descriptive approach assuming that individuals dislike travel time variability, but it does not purport to explain why.

- Scheduling approach: this approach is strongly linked to the departure time choice (or trip scheduling) studies. The model developed by Small (1982) is based on choices under certainty. Noland and Small (1995), developed Small's scheduling model to analyse and try to understand the choice of departure time under uncertainty, adding the probability distribution of travel time. Given travel time variability, travel time (T) is uncertain with a distribution dependent on the departure time (t_h)(Bates et al. (2001)).

$$E(U(t_h)) = \alpha E(T(t_h)) + \beta E(SDE(t_h)) + \gamma E(SDL(t_h)) + \theta P_L(t_h) \tag{2.15}$$

According to Bates et al. (2001), the scheduling model and the mean-variance model can be approximated under certain conditions:

  - travel time distribution is independent of departure time
  - there is no lateness penalty
  - departure time is continuous
  - regular congestion is independent of departure time

A recent work by Fosgerau and Karlström (2010) mathematically demonstrated the previous equivalence statement by Bates et al. (2001).

- Mean Lateness: this approach is commonly used for measuring reliability for passenger rail transport in the UK. Travel unreliability is measured by the mean lateness -defined as the difference between schedule departure and actual departure (lateness at boarding) and time between schedule arrival and actual arrival (lateness at destination). The first formulation considering this approach was made by the Association of Train Operating Companies (ATOC, 2005) :

$$E(U) = \gamma_1 SchedT + \gamma_2 L^+ \tag{2.16}$$

where $SchedT$ is the scheduled travel time and L^+ is the mean lateness at the destination train station. Batley and Ibáñez (2012) expanded this framework to include lateness at boarding (B^+) and a train fare in the expected utility function.

2.5 Conclusion

This review of the economic literature on congestion was based on the example of road transport. We considered that the main theoretical conceptual models of individual behaviour, and particularly equilibrium models, have been developed extensively in the road sector. This theoretical development in the road sector allowed us to describe all the main significant future concepts of our research with a certain homogeneity.

In contrast to the literature dealing with road transport, research in the other transportation modes has not essentially dealt with the issue of trip scheduling and passenger behaviour. It has focused on the mechanisms generating congestion and their monetary costs, as was shown in the introduction, or on the possibility of using pricing to deal with congestion externalities, as in the road sector. For example, in the railway sector, capacity shortage has traditionally been considered as the inability of a train operator to obtain the desired train path (scarcity). However, this perception of capacity seems restrictive. A lack of capacity can occur before scarcity, as unexpected transmitted delays are positive in relation to traffic density (congestion).

The aim of this PhD research is to contribute to the development of the economic analysis of rail capacity constraints. This research applies the theoretical concepts developed in other transport modes, adding the particularities associated with rail transport.

Considering the lessons from an engineering perspective (chapter 1) and from the theoretical economic viewpoint (chapter 2), chapters 3 and 4 develop an original user's travel cost function for rail passengers, and a supply-demand equilibrium model, respectively. The main particularity in rail transport is that the users of scheduled services cannot choose their departure time freely, but are constrained to the departure times of the service. Consequently, it is important to understand how frequencies affect the user's travel cost function.

Chapter 3

Rail Capacity Constraints in the Consumer Generalised Cost Function

Contents

3.1 Introduction

Chapter 3 proposes a new user generalised cost function approach. The microeconomic model described in this chapter incorporates a theoretical measure of the value of frequency for transport services with planned timetables (trains, buses and planes).

In this research, travel time cost for users using transport services with fixed timetables is decomposed into three components. Firstly, a **planned travel time** which corresponds to the announced travel time for a trip. Secondly, due to imposed fixed timetables, travellers suffer an **expected schedule delay cost**, travelling either earlier or later than they would like to. Finally, in case of unexpected incidents and in a highly traffic density situation, users can bear a **random delay cost**, if the transport arrives later than scheduled.

Usually and as exposed in the previous chapter, the trip scheduling preferences concept is used to analyse congestion situations, where users apply a trade-off between travel time and arrival-time scheduling preferences (Vickrey, 1969; Small, 1982; Arnott et al., 1993). Nevertheless, in this PhD dissertation, scheduling preferences are used to analyse a situation in which the infrastructure manager looks for an optimal frequency knowing that there is a trade-off between the expected schedule cost for users - high frequency means fewer scheduling costs - and the random delay cost for users - high frequency would facilitate delay propagation.

On the one hand, previous papers have focused on the impact of frequency on expected schedule delay costs, considering that travel time is deterministic. Mohring (1972) studied the impact of the number of users on frequency and fares for public buses, by considering that if both the number of travellers and frequency increase, the waiting cost for users will diminish. Jansson (1993) sought an optimal price and frequency by considering scheduling cost for users who either plan or do not plan their trip. De Palma and Lindsey (2001) investigated the optimal scheduling of a given number of public transport vehicles in a single line network. Lastly, Fosgerau (2009) presented a trade-off between scheduling costs and waiting time in services with short and long headways.

On the other hand, and as detailed in chapter 2, previous research has also dealt with the random delay cost formalisation. The initial scheduling model (Small, 1982) was further developed considering stochastic travel time, e.g: Noland and Small (1995), Bates et al. (2001), Fosgerau and Karlström (2010), Fosgerau and Engelson (2011) .

However, until now, the cost-benefit analysis of expected schedule delay costs and

random delay costs has been tackled independently in transport with timetables planned without considering that a trade-off exists between both parameters.

The generalised cost function for train users set out in this research proposes a new combined vision of these two concepts, by applying the empirical trade-off between schedule delay costs and random delay costs, demonstrated in chapter 1.

The issue of the trade-off between passengers' schedule delay cost and random delay cost has also been considered recently by Lin and Zhang (2016) for air transport. However, the generalised cost function developed in this PhD research is more detailed analytically, and in particular considers a sophisticated random delay cost function.

3.2 Theoretical Model

The methods described in chapter 1 demonstrate that the current construction process of a train timetable (in France and elsewhere) takes into account the logic of quality for the service supplied. All other things being equal, there is a trade-off between train path robustness (the capacity to recover from an incident) and travel time, and between train diagram robustness and rail capacity. Nevertheless, the level of these trade-offs has not as yet been objectivised. The goal of this chapter is to establish a generalised cost function that reflects all the trade-offs concerning capacity constraints for consumers.

The main objective is to take into account the costs for users associated with rail capacity constraints. Consequently, user's travel cost will be composed of an expected schedule delay cost (due to the difference between preferred arrival time for users and timetabled arrival time) and of a random delay cost, which stands for the increased cost of delay as a function of traffic density. If traffic density is high, headways between trains will be smaller and the delay snowball effect will be higher.

3.2.1 Model set-up

We consider a simplified network with a double track line with homogeneous traffic between two train stations.

For the general specification of the model, the following assumptions are made:

- As rail transport is a scheduled transport mode, the infrastructure manager establishes a frequency f (number of trains/unit time) in advance between two cities ($f > 0$).

- Demand N is uniformly distributed throughout the unit time T ($T > 0$). We consider $N = zT$, where z is the number of users per unit time.

The aim of the model is to identify the optimal frequency which minimises the monetised time cost function for passengers, bearing in mind all the adjustments described in chapter 1. We assume that user's monetised time cost function GC_0 is a linear function of planned travel time cost C_T, expected schedule delay cost C_E and random delay cost C_R:

$$GC_0 = C_T + C_E + C_R \tag{3.1}$$

considering that $C_T, C_E, C_R > 0$.

In addition to the usual planned travel time cost, two types of schedule cost coexist in this modelling:

- An expected schedule delay cost, based on the discreteness of transport service timetables.
- A random delay cost that depends on unexpected events on the network and traffic conditions.

The next section describes the methodology employed to model each component of the cost function 3.1. We consider that the planned travel time is given by the equation $C_T = \alpha T$ where α is the value of travel time ($\alpha > 0$).

3.2.2 Expected schedule delay cost

Trains are equally spaced during the unit of time considered. The difference between the preferred arrival time by users and the arrival time fixed by the infrastructure manager represents the expected schedule cost for users. Arrival time is based on the timetable announced and does not consider random delays. In this dissertation, we do not take into account that travellers may anticipate the possibility of delays when choosing their optimal arrival time as did Tseng and Verhoef (2008) and Tseng et al. (2012). An extended analysis considering this possibility is developed in appendix B.

According to the concept underlying the location models of Hotelling (1929) and Salop (1979), we consider that each consumer has a most preferred arrival time t^* with $t^* \varepsilon [0, T]$. In a transport mode where frequency f are discrete and fixed in advance by

the infrastructure manager, passengers must adjust their most preferred arrival time to the timetabled arrival times. This difference between the preferred and the timetable arrival times induces a disutility for each passenger. We assume that 0 and T are the most preferred extremes. The value T is the operating time interval.

We consider that a passenger's most preferred arrival time is t^* (Figure 3.1). As no train arrives exactly at time t^*, the passenger chooses between taking a train arriving before (i.e $T1$) – thus being ahead of time at their destination - or a train arriving after (i.e $T2$) – thus being late. H is the effective interval time between two trains ($H^*\varepsilon[0;T]$).

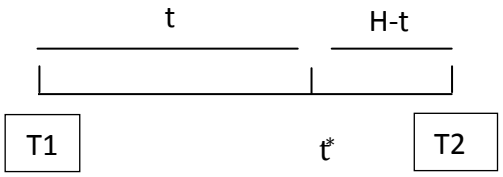

Figure 3.1: Preferred travel time

According to the road congestion literature, (Noland and Small, 1995; Arnott et al., 1990), we consider that arriving at a time different from the preferred arrival time represents a cost for travellers. However, the phenomenon is different here because users rely on timetabled departure-arrival times. This contrasts with travel by car, where users can depart at any time. On car trips, passengers apply a trade-off between travel time (trying to avoid peak-period congestion) and schedule delay for trip timing decisions. However, most users of scheduled transport are subject to an expected schedule delay even if the transit system is reliable and keeps perfectly to the timetable (De Palma and Lindsey, 2001).

In this chapter, we delimit the cost due to this time imbalance as the "expected schedule delay cost", with the following definition:

- If a passenger decides to arrive at T_1, they should leave their other activities early (wake up early, leave work early, etc.). Moreover, they will arrive before their preferred time at the station. The cost is βt, where β is the schedule delay cost of arriving early (before t^*).

- If they decide to arrive at T_2, they could wake up later or stay at home longer, but will arrive later than the preferred time at their destination, with a cost $\gamma(H-t)$, where γ is the schedule delay cost of arriving late (after t^*).

$$C_{E(Early)} = \beta t \tag{3.2}$$

$$C_{E(Late)} = \gamma(H - t) \tag{3.3}$$

In order to calculate the disutility generated, we must first estimate the location t_i of passenger indifference between both alternatives. Figure 3.2 represents the utility function as a function of the preferred arrival time. This utility function is always equal to or higher than 0. Of course, it is 0 if the preferred arrival time coincides with the actual arrival of a train.

$$t_i = T1 + \frac{H\gamma}{\beta + \gamma} \tag{3.4}$$

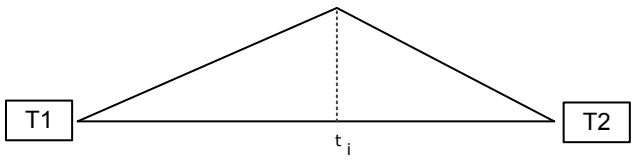

Figure 3.2: Scheduling delay cost

Once the time t_i has been determined, the passengers compare their preferred travel time t^* with t_i and decides which train to choose. If $t^* < t_i$ they would choose the previous train. Otherwise, they would choose the next train (with a utility $U(t^*)$) . Given his decision, they assume the disutility associated with the difference between t^* and the train arrival chosen. As shown before, this equals the slope of the straight lines defined by equations 3.2 and 3.3, respectively. As we have considered that the demand is uniformly distributed throughout the interval T, the demand for the interval H equals kH. The total disutility for passengers in the interval H equals the total area of the triangle[1].

$$Base = zH = \frac{1}{f}z \tag{3.5}$$

$$Height = \frac{H\beta\gamma}{\gamma + \beta} \tag{3.6}$$

$$C_{E(Total)} = C_{E(Early)} + C_{E(Late)} = H^2 \frac{\beta\gamma}{2(\gamma + \beta)} z \tag{3.7}$$

[1]It is assumed that individuals are identical except for their desired travel times.

If we consider that the number of travellers k per train is

$$k = zH \tag{3.8}$$

Consequently, the average consumer's cost for time adjustment can be represented by a function of the form:

$$C_{E(Average)} = H\frac{\beta\gamma}{2(\gamma+\beta)} \tag{3.9}$$

Equation 3.9[2] shows that the expected schedule delay cost increases when the headway H between two train increases, which means that it decreases with frequency f, and equation 3.9 can be rewritten as:

$$C_{E(Average)} = \frac{1}{f}\frac{\beta\gamma}{2(\gamma+\beta)} \tag{3.10}$$

This relationship underlines an identified effect in transport economics when timetables are scheduled in advance, known as the " Mohring effect".

The relationship between frequency and waiting time costs has its origins in an analysis performed for public transport, and particularly for buses (Mohring, 1972). In the context of public transport, users are considered to arrive randomly. Their waiting time cost at the bus stop depends on bus service frequency. A higher level of demand in a given geographical area will generate a reduction in the total travel cost for users, due to the increment on frequency and the reduction on their travel time. This effect is known as the Mohring effect.

In rail transport, the timetable is fixed in advance and users are not assumed to arrive randomly. Therefore, a variation of frequency does not mean a variation of waiting time cost at the station. By contrast, a change in frequency and timetable involves a variation of the expected schedule delay cost. According to the Mohring effect, in a context of services scheduled in advance, a change of frequency incurs a variation on the expected schedule delay cost.

In this section, it has been considered that users experiment with an expected schedule delay cost when they decide to travel due to the impossibility of perfectly adjusting their preferred arrival time. This expected schedule delay is independent of travel time reliability: random delays costs are considered in the next section.

[2]This result is equivalent to proposition 2 of De Palma and Lindsey (2001) and Fosgerau (2009).

3.2.3 Random delay cost

As stated in the introduction, in some cases, stochastic delays can increase the schedule travel time and delay propagation will depend on traffic density.

As described before, the infrastructure manager considers several margins to mitigate the risk of delays when they design the train diagram, but this does not mean the absence of delays in the network: several stochastic delays still exist due to unexpected events (asset failures, weather conditions, passenger behaviour, etc.).

Robust reliability indicators are needed to link train reliability to capacity utilisation. Carey (1999) presented an insightful analysis of the mechanism underlying delays. He considered two types of delays: exogenous or primary delays and knock-on or secondary delays. Exogenous delays are due to events such as the breakdown or failure of equipment or infrastructure, delays in passenger boarding, lateness of operations or crews, etc. Generally, exogenous delays are not due to scheduling issues. Conversely, knock-on delays are due to exogenous delays and their interdependence in the schedule. Under high utilisation, a delayed train can cause delays to several other trains over a large area and a long period of time. As stated in the introduction, the relationship between intensive usage and a degradation in the quality of service, as capacity remains fixed is known as "congestion". Knock-on or secondary delays can be reduced by scheduling, for example by giving more headway to trains prone to exogenous delays.

Like Villemeur et al. (2015), we do not consider that primary delays depend on the pattern of flows. Our intuition is that the probability of a primary delay is given and independent of the number of flows (technical problems or human errors are not a function of the number of trains running in our link). The recovery time T_{extra} considered in the scheduling process can allow for recovery from an incident in some cases.

Nevertheless, in contrast to Villemeur et al. (2015), we consider that the model's specifications should also reflect congestion issues. Indeed, the origin and probability of an incident are effectively independent of the number of trains; however, the consequences of these events are strongly linked to them (trains/unit time). When a train track is used intensively, an additional train path increases the consequences of delays, due to a reduction in the capacity to recover from an incident. This means that when traffic is high, the probability of spreading delays is higher and thus their total effects are greater. As with airports, rail congestion exhibits a cascade-type effect: a single delay may generate an impact which accumulates over the subsequent trains.

As described in chapter 1, in order to control delay propagation, a buffer time H_{extra} between trains is introduced in the scheduling process. High capacity consumption results in higher risks of consecutive delays. If there is enough buffer time between two trains, small delays will not affect the successive train(s). When a primary delay propagates to another train, a secondary delay can arise. In line with Landex (2008), the description of propagation delay in this model assumes a double track line with homogeneous one way operation on each track (meaning that both the speed and the buffer time are constant).

Buffer time between two trains can be expressed as the difference between H which is the effective headway between two consecutive trains and H_{min} which is the minimum technical headway ($H, H_{min}, H_{extra} > 0$)

$$H_{extra} = H - H_{min} \tag{3.11}$$

Given equation 3.11 we can write the maximal capacity/frequency of the line as [3] :

$$f_{max} = \frac{1}{H_{min}} \tag{3.12}$$

And the frequency f :

$$f = \frac{1}{H} = \frac{1}{H_{min} + H_{extra}} \tag{3.13}$$

The delay function considered in this research combines the two previous studies (Villemeur et al., 2015; Landex, 2008).

We consider a stochastic incident ε ($\varepsilon > 0$), independent of traffic flows. The amount of delay for the first train is $d_{1,i}$

If $d_{1,i} > 0$, the primary delay can be propagated to the subsequent trains, depending on the level of buffer time between trains. The amount of delay propagation, or consecutive delay for the following train $d_{2,c}$, can be calculated as:

$$d_{2,c} = \begin{cases} d_{1,i} - H_{extra} & \text{si } \mathrm{H}_{extra} < d_{1,i} \quad (3.14) \\ 0 & \text{sinon} \quad (3.15) \end{cases}$$

If the buffer time H_{extra} is longer than or equal to the delay $d_{1,i}$, the delay will not lead to a consecutive delay of the succeeding train, $d_{2,c}$ will then be less than or equal to

[3]Reminder: We consider homogeneous and uniformly distributed traffic.

zero. Formula 3.14 can be generalised to calculate the consecutive delay for any of the following trains where there are no longer initial delays

$$d_{j+1,c} = d_{1,i} - xH_{extra} \tag{3.16}$$

In equation 3.16, x is the number of trains affected by consecutive delays. By setting the consecutive delay $d_{t+1,c}$ equal to zero (meaning that the last train will have no consecutive delay), it is possible to calculate the number of trains needed before the trains run on time again. A train is either delayed or on time, therefore, the decimal numbers must be truncated:

$$x = \left\lfloor \frac{d_{1,i}}{H_{extra}} \right\rfloor \tag{3.17}$$

And the total number of delayed trains is equal to $X = x + 1$

Knowing the number of trains x having consecutive delays, it is possible to calculate the total delay, which is equal to the sum of consecutive delays and the initial delay:

$$\sum d = d_{1,i} + d_{2,c} + d_{3,c} + d_{4,c} + \cdots + d_{x+1,c} = d_{1,i} + \sum_{k=1}^{x+1} d_{x,c} \tag{3.18}$$

By combining formulas 3.16 and 3.18, the total delay can be rewritten as:

$$\sum d = d_{1,i} + d_{1,i} - H_{extra} + d_{1,i} - 2H_{extra} + \cdots + d_{1,i} - xH_{extra} = (x+1)d_{1,i} - \frac{x}{2}(x+1)H_{extra} \tag{3.19}$$

By combining formulas 3.17 and 3.19, the total delay can be calculated based on the initial delay ($d_{1,i}$) and the buffer time (H_{extra}):

$$\sum d = \left(\left\lfloor \frac{d_{1,i}}{H_{extra}} \right\rfloor + 1 \right) d_{1,i} - \frac{1}{2} \left\lfloor \frac{d_{1,i}}{H_{extra}} \right\rfloor \left(\left\lfloor \frac{d_{1,i}}{H_{extra}} \right\rfloor + 1 \right) H_{extra} \tag{3.20}$$

As a delay is a random variable not known with certainty in advance, it will henceforth be convenient to use the expected delay as the formulation for random schedule delays $E(\sum d)$.

Defining $E(\sum d)$ by considering the number of trains having consecutive delays (Formula 3.17) is very complex from the mathematical viewpoint and some approximations must be taken into account:

$$X \simeq \frac{d_{1,i}}{H_{extra}} + 1 \tag{3.21}$$

$$X(X-1) \simeq \left(\frac{d_{1,i}}{H_{extra}}\right)^2 \tag{3.22}$$

$$\sum d = \frac{d_{1,i}^2}{2H_{extra}} + d_{1,i} \tag{3.23}$$

Considering the previous approximations, $E(\sum d)$ can be specified as:

$$E\left(\sum d\right) = \frac{E(d_{1,i}^2)}{2H_{extra}} + E(d_{1,i}) \tag{3.24}$$

and based on the Köning Huygens theorem,

$$E\left(\sum d\right) = \frac{\mu_{d_{1,i}}^2 + \sigma_{d_{1,i}}^2}{2H_{extra}} + \mu_{d_{1,i}} \tag{3.25}$$

where μ represents the average initial delay and σ the standard deviation of the initial delay.

This formula is based on the propagation delay function described in Landex (2008) and has been completed by considering that a delay is a random variable. From our viewpoint, this approach provides a comprehensive notion and completes the functions provided by Villemeur et al. (2015) and Landex (2008).

The total delay function 3.25 reveals that adding a buffer time (H_{extra}) decreases the total delay. By contrast it limits total capacity and thus the frequency supplied.

Delays logically increase travel time for users. Delays as unexpected events present higher costs for passengers than costs related to schedule travel times. We consider a random delay time cost δ .

Taking this into account, we can rewrite equation 3.25 as the total delay cost for passengers as:

$$C_{R(Total)} = \delta \left[\frac{\mu_{d_{1,i}}^2 + \sigma_{d_{1,i}}^2}{2H_{extra}} + \mu_{d_{1,i}}\right] \tag{3.26}$$

Combining functions 3.13 and 3.26, it is feasible to express the average delay cost function of the frequency f[4]:

[4]The average delay for a train is equivalent to the average delay for consumers

$$C_{R(Average)} = \delta \left[\frac{\mu_{d_{1,i}}^2 + \sigma_{d_{1,i}}^2}{2(\frac{1}{f} - \frac{1}{f_{max}})} + \mu_{d_{1,i}} \right] \tag{3.27}$$

3.3 The Issue of Passenger Optimisation: Analytical Solution

Considering all the assumptions described in the previous section, it is possible to define the optimal frequency that minimises the generalised cost function for users:

The minimisation problem in views to obtaining the optimal frequency, can be written as :

$$Min_f GC_0 = \alpha T + \frac{1}{f} \frac{\beta\gamma}{2(\gamma+\beta)} + \delta \left[\frac{\mu_{d_{1,i}}^2 + \sigma_{d_{1,i}}^2}{2(\frac{1}{f} - \frac{1}{f_{max}})} + \mu_{d_{1,i}} \right] \tag{3.28}$$

We consider a benevolent infrastructure manager which wishes to maximise the net utility for passengers from travelling by train between two cities, by considering the associated costs: planned travel time costs, expected schedule delay costs and random delay costs. The infrastructure manager seeks an optimal frequency, knowing that, all other things being equal, high frequency mean fewer expected schedule delay costs (second right hand term) but, correspondingly, more expected random delays costs (last right hand term).

The first-order necessary conditions are:

$$\frac{\partial GC_0}{\partial f} = 0 = \frac{\delta \left({\sigma_{d_{1,i}}}^2 + {\mu_{d_{1,i}}}^2 \right)}{2\, f^2 \left(\frac{1}{f} - \frac{1}{f_{max}} \right)^2} - \frac{\beta\, \gamma}{2\, f^2\, (\gamma+\beta)} \tag{3.29}$$

At equilibrium, the infrastructure manager would choose an optimal f^* from the consumer's perspective. This level of frequency ensures that the marginal cost of expected schedule delay cost (Mohring effect) adjustments equals the marginal random delay cost (congestion effect).

$$f^* = -\frac{{f_{max}}^2 \sqrt{{\sigma_{d_{1,i}}}^2 + {\mu_{d_{1,i}}}^2}\, \sqrt{\beta\, \delta\, \gamma^2 + \beta^2\, \delta\, \gamma} + \beta\, f_{max}\, \gamma}{\left(\delta\, {f_{max}}^2\, {\sigma_{d_{1,i}}}^2 + \delta\, {f_{max}}^2\, {\mu_{d_{1,i}}}^2 \right)\, (\gamma+\beta) - \beta\gamma} \tag{3.30}$$

Each additional f decreases the marginal scheduling cost and at the same time increases the marginal delay cost .

Once solved for the optimal frequency (f^*) and following equation 3.11 it is possible to calculate the optimal buffer time H^*_{extra} :

$$H^*_{extra} = \frac{1}{f^*} - H_{min} \tag{3.31}$$

The second order condition is also verified [5]:

$$\frac{\partial^2 Gc_0}{\partial f^2} = \frac{\beta\,\gamma}{f^3\,(\gamma+\beta)} - \frac{\delta\,\left({\sigma_{d_{1,i}}}^2 + {\mu_{d_{1,i}}}^2\right)}{f^3\left(\frac{1}{f} - \frac{1}{f_{max}}\right)^2} + \frac{\delta\,\left({\sigma_{d_{1,i}}}^2 + {\mu_{d_{1,i}}}^2\right)}{f^4\left(\frac{1}{f} - \frac{1}{f_{max}}\right)^3} < 0 \tag{3.32}$$

The monetised time cost function for users $GC_0(f)$ can be represented graphically. At the beginning, the average $GC_0(f)$ slopes downwards because of significant positive interactions between train frequency: if the number of frequency increases, the expected schedule delay cost will decrease (Mohring Effect).

In contrast, on the second part of the curve we observe an upward-sloping trend: if the number of frequency increase, the random delay cost (congestion effect), depending positively on traffic density, will increase.

The average cost curve represents the function on which individual train operator companies (TOC's) base their frequency choice demands.

The marginal cost $GC_0(f)$ curve is obtained by:

$$MC = \frac{\partial f GC_0(f)}{\partial f} = GC_0(f) + f\frac{\partial GC_0(f)}{\partial f} \tag{3.33}$$

The first term of equation 3.33 represents the average cost and the second term is the marginal external cost of an additional train.

As can be seen in figure 3.3, the marginal external cost is negative at the beginning of the MC curve. A negative marginal external cost means a positive externality: an additional frequency decreases the social cost of travelling for all users. Indeed, increasing frequency decreases the expected schedule delay cost associated with the preferred travel time for the other users. In the second part of the figure, the marginal cost curve lies above the average cost. This means that there are negative externalities: the additional random delay that a new frequency imposes on the others, which is not taken into account by the last train assigned to travel.

[5]The second order condition has been verified using numerical values.

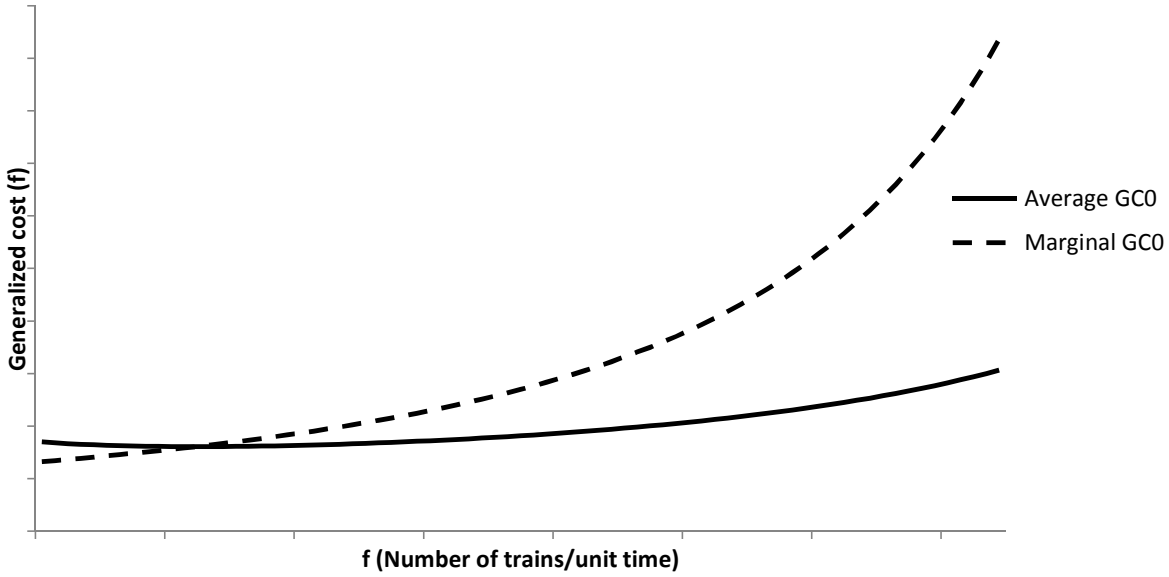

Figure 3.3: Average and marginal monetised time cost function

3.4 The Issue of Passenger Optimisation: a Few Graphical Illustrations

Up to now, the optimal frequency from the consumer's perspective has been defined analytically. In order to further illustrate the properties of the model, this section will display several numerical results.

3.4.1 Comparative statics

In view to obtaining better understanding of the relationship between the optimal frequency and the other parameters of the model, the following figures illustrate how a variation on their numerical values affects the optimal frequency. Taking the available data into account, it is not the purpose of this section to precisely describe a real-life rail system. The parameters presented therefore do not have to correspond to real life values.

In the next figures we represent the values of the multipliers associated with each time parameter. Since utility is linear in all the parameters and for simplicity in the interpretation results, we have to consider the Value of Travel Time Savings (VTTS) as

equal to 1. Consequently, the value of the multipliers is equal to the total value time of each parameter.[6]

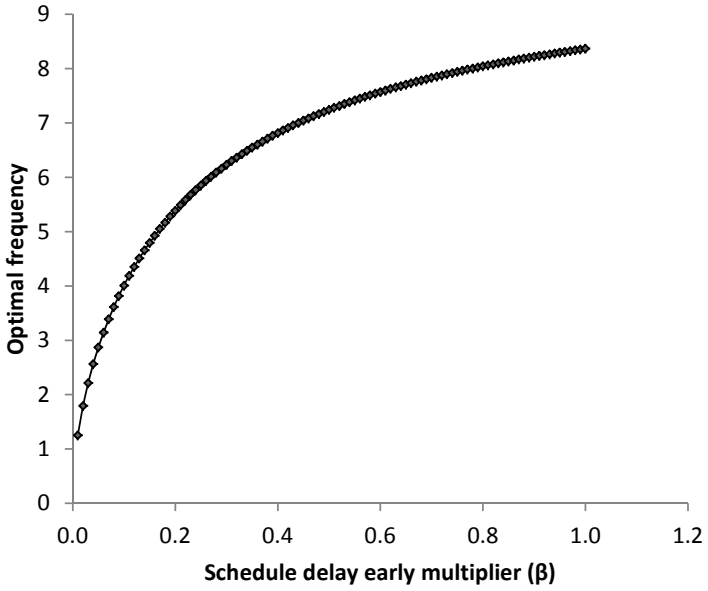

Figure 3.4: Influence of schedule delay early multiplier on optimal frequency

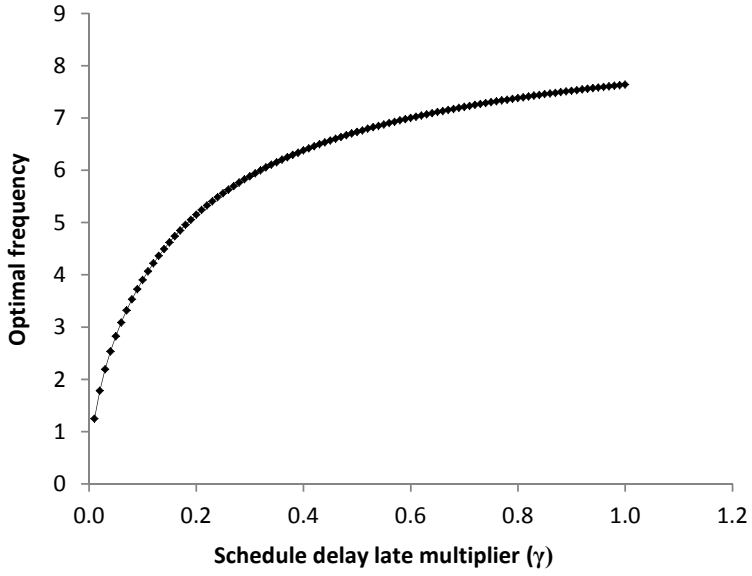

Figure 3.5: Influence of schedule delay late multiplier on optimal frequency

[6]The multiplier values used in the simulations are detailed in the appendix C

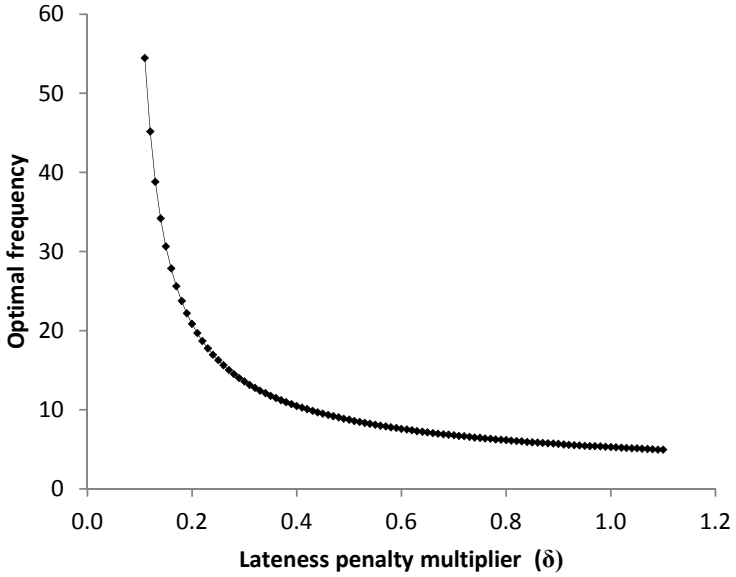

Figure 3.6: Influence of lateness penalty multiplier on optimal frequency

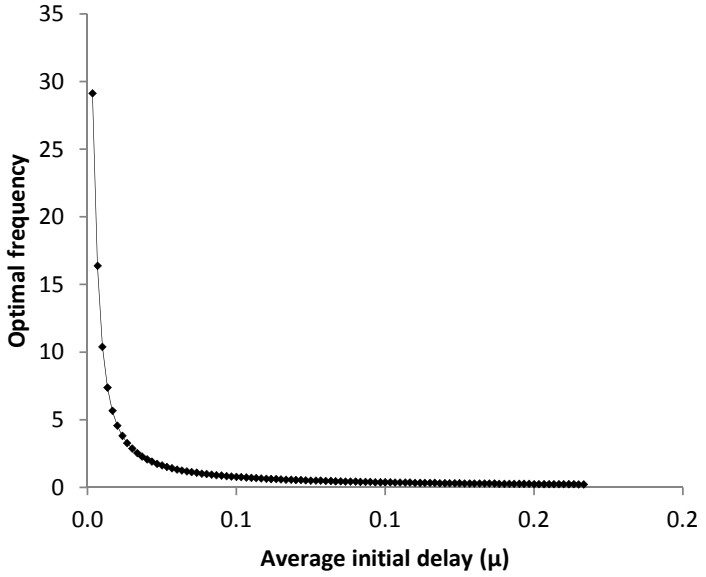

Figure 3.7: Influence of average initial delay on optimal frequency

As can be seen in figure 3.4 and figure 3.5, expected schedule delay costs are positively correlated with optimal frequency for passengers. If the cost of not having frequency at the desired travel times were high for passengers, they would prefer to have higher frequency for their trips.

In contrast, in figure 3.6 we note a negative relationship between the lateness penalty

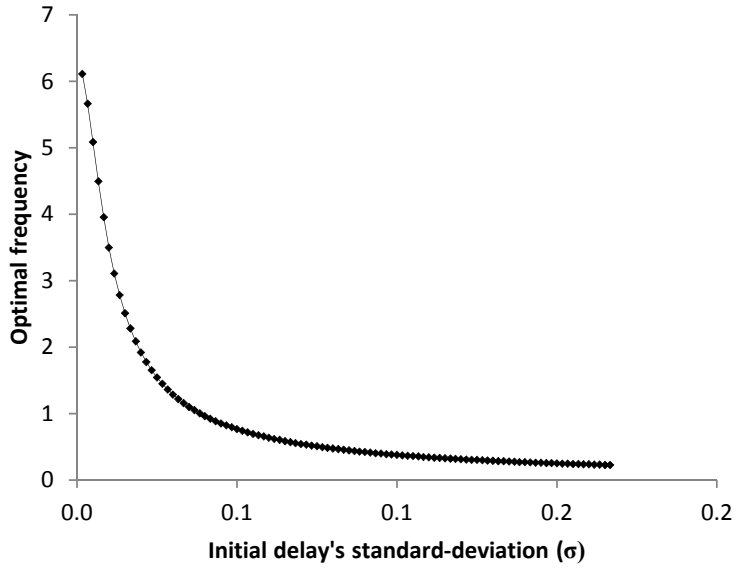

Figure 3.8: Influence of initial delay's standard deviation on optimal frequency

multiplier (associated to the random delay cost) and the optimal frequency. If the penalty of arriving late, after the scheduled arrival time, were high, passengers would prefer to have lower frequency. Indeed, passengers know that if frequency is high, the probability of being late is higher too.

Figures 3.7 and 3.8 illustrate a negative relationship between the initial delay and the optimal frequency for passengers. If the initial delay is considerable, considering the average delay or the standard deviation, it would be more difficult for the network to recover from an incident and delay propagation would snowball.

3.4.2 Sensitivity analysis

As stated above, current available data do not allow us to determine a feasible interval of optimal frequency for each line easily and with precision. For example, empirical estimations concerning the expected schedule delay multipliers are rare for scheduled public transport. Apart from the studies by Nuzzolo and Russo (1998) and De Palma and Fontan (2001), in which the values concern public transports, no other research on the subject has been identified.

In this context, the main idea of this section is to analyse whether parameters exist that significantly affect the relationship between optimal frequency and the other parameters. The idea is then to identify if the uncertainty around certain parameters (as β, γ) greatly affects the results and to determine the most "sensitive" ones. Once the latter have been determined, it will be easier to identify the objectives of future empirical research on this topic.

The aim of this section is to analyse how the relationship between optimal frequency and one parameter is affected by changes to the other parameters. To illustrate the use of sensitivity analysis, we have considered how the relationship between optimal frequency and the schedule delay early multiplier evolve by making changes to the other parameters . The objective is to determine whether or not a change made to one parameter significantly affects the previous relationship.

The interval of values chosen for each parameter is based on reasonable values, considering available economic literature and industrial data[7]. Nevertheless, as previously stated, they do not represent real-life values and the following figures do not allow making formal recommendations on optimal frequency values.

Figure 3.9 determines if the relationship between the schedule delay early multiplier (β) and the optimal frequency is strongly affected or not by a change in the schedule delay late multiplier (γ).

Firstly, we observe that the higher γ is, the higher the optimal frequency. That seems consistent with the previous comments on figures 3.4 and 3.5. Secondly, we observe that the sensitivity due to the variation of γ, depends on the initial value of β. Indeed, in the first part of the figure, the optimal frequency is relatively independent of the variation of γ. Nevertheless, at the end of the curve, modifying the value of γ could double the optimal frequency, which is a considerable difference.

[7]The parameters values used in the simulations are detailed in the appendix C

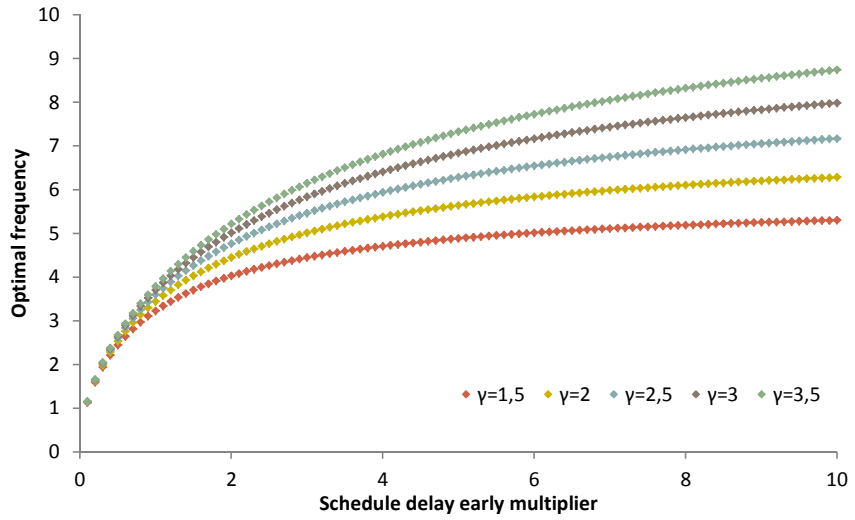

Figure 3.9: A change in the value of schedule delay late multiplier

Although the data on β is not extensive, the few empirical values available estimate β between the interval $[1.63 - 2.92]$ (Nuzzolo and Russo, 1998; De Palma and Fontan, 2001). Considering these values, it seems pertinent to focus on the first part of the figure. In this context, the optimal frequency variations are relatively independent of the value of γ.

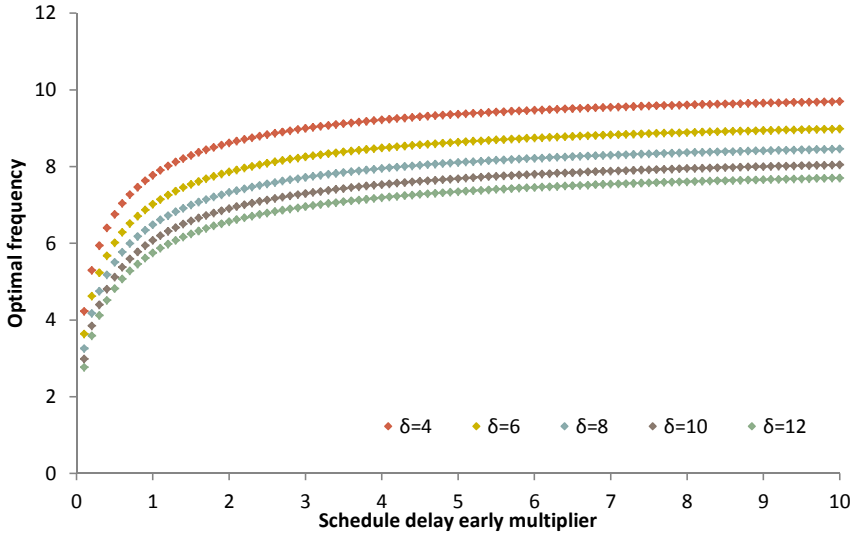

Figure 3.10: A change in the value of the lateness penalty multiplier.

Figure 3.10 determines whether the relationship between the schedule delay early multiplier (β) and optimal frequency is strongly affected or not by the change in the

lateness penalty multiplier (δ).

Firstly, we observe that the higher δ is, the lower the optimal frequency. That seems consistent with the previous comments on figure 3.6. In the second instance, we observe that a variation of δ, does not significantly change the optimal frequency.

Figures 3.11 and 3.12 determine if the relationship between the schedule delay early multiplier (β) and the optimal frequency is substantially affected or not by the change in the average or standard deviation of the initial delay (μ and σ).

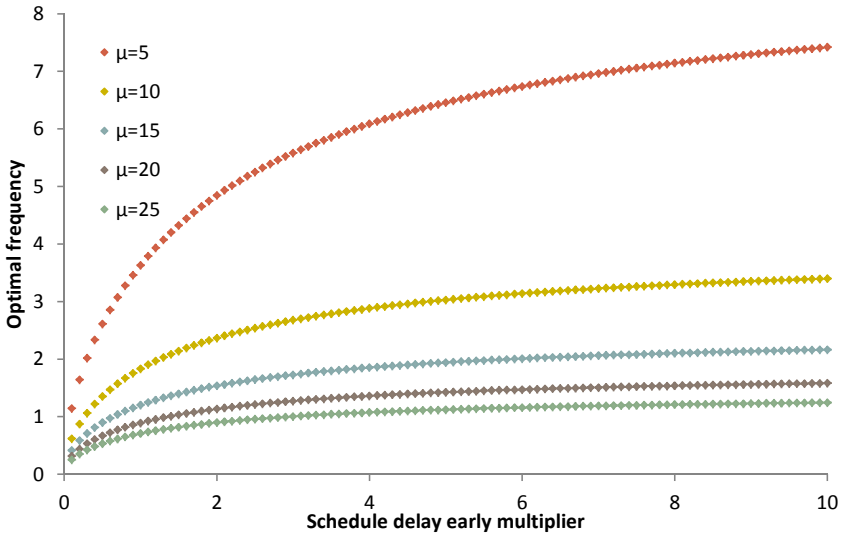

Figure 3.11: A change in the average initial delay value.

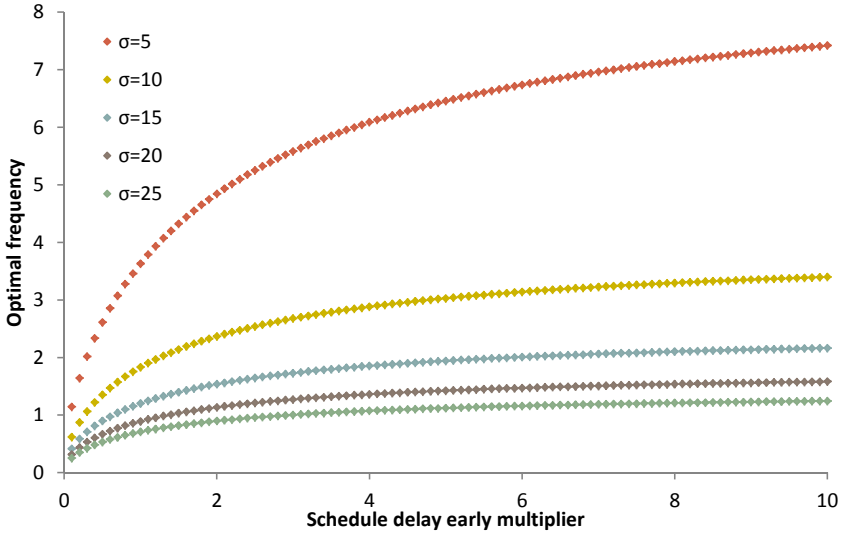

Figure 3.12: A change in the standard deviation of the initial delay value.

The results of these figures appear consistent with those of figures 3.7 and 3.8: the higher μ and/or σ are, the lower the optimal frequency. Furthermore, we observe that a variation in μ and/or σ noticeably affects the optimal frequency. We can observe that considering an average initial delay (μ) of 5 or 10 minutes, halves the optimal frequency. Although the optimal frequency presented cannot be interpreted as real-life values, this means that a lower variation in the average initial delay (or in the standard deviation of the initial delay) implies a considerable effect on optimal frequency, since both are "sensitive" parameters in the model.

Sensitivity analysis allows characterising the relationship between variables and determining the most influencing parameters. The previous figures show that the characteristics of the initial delay (μ and σ) seem, in principle, to have the highest impact on the choice of optimal frequency. It appears essential to have in-depth knowledge on the initial delay parameters (μ and σ) in order to determine the optimal frequency.

From the practical viewpoint, the average and standard deviation of the initial delay are data available to the infrastructure manager. These data are not calculated systematically for each line, but the infrastructure manager has the original data that enables them to calculate this variable precisely.

Although the previous figures show that precise knowledge of the expected schedule delay multipliers (β and γ) is not decisive for choosing the optimal frequency, it could be interesting for the infrastructure manager to carry out in-depth empirical research to determine its value. In addition to formulating the calculus of optimal frequency, better knowledge of these parameters would allow the infrastructure manager to better evaluate certain infrastructure projects from a socio-economic perspective.

In contrast with the expected schedule delay multipliers, there are numerous studies concerning he lateness penalty multiplier. For example, an interesting study was developed along these lines by the French infrastructure manager to better understand and measure the lateness penalty multiplier in the CBA analysis in France some years ago (Guiraud et al., 2014). In Great Britain, Batley et al. (2011) have also estimated the elasticity of demand for rail with respect to changes in response to changes in service performance (lateness and reliability).

3.5 Conclusion

The description of the current graphic timetable construction process in chapter 1 highlighted that, all other things being equal, a relationship exists between traffic volume, delays and train path scarcity. From the consumer perspective, the rail capacity constraint can be expressed in two complementary but not exclusive ways. On the one hand, there is a schedule delay effect or Mohring effect (the impossibility of travel at the preferred travel time) and on the other hand, a congestion effect (relationship between traffic and delays).

The microeconomic model proposed in this dissertation reveals that there is an implicit trade-off between the schedule delay effect or Mohring effect and the congestion effect when the IM supplies frequency in the current graphic timetable construction process. However, the link between these two variables has not been formalised and measured until now. The aim of this formalisation is to define the optimal frequency f^* which minimises the generalised user cost function, considering all the relationships specified.

Optimal frequency depends on several parameters and their analysis allows illustrating the properties of the model. This chapter proposed the first step in the analysis of these relationships, but the results should be developed in further research. The results of the theoretical model could be evaluated empirically using a calibration model. The aim of the latter is to determine a feasible interval of optimal frequency, taking into account the variability of the others parameters.

As shown by the sensitivity analysis, some parameters have a greater impact on the optimal frequency calculation than others. As stated previously, it is essential in the first phase to have an accurate understanding of the initial delay characteristics (μ and σ) of each line. Secondly, if infrastructure managers wish to have better understanding of passenger behavioural parameters such as β and γ, it will be necessary to develop specific empirical research and surveys.

The calibration of the parameters should be based on the empirical values available in the academic literature and in future empirical developments (schedule delay multipliers, lateness penalty multiplier and time values) and in infrastructure management databases (average initial delay and standard deviation of initial delay).

These more extended analyses could help identifying the value of frequency and buffer times, which, as a function of certain values of the other parameters, minimise the monetised time cost function for users.

Until now, both in the literature and in practice, the cost-benefit trade-off of frequency (schedule delay effect and congestion effect) has been examined independently in planned transport services. The generalised cost function described in this chapter offers a new perspective, which considers both effects simultaneously and proposes a detailed formalisation of both concepts.

Chapter 4

Supply-Demand Equilibriums

Contents

4.1 Introduction

Now that the detailed generalised cost function for train users has been determined in the previous chapter, chapter 4 seeks to build an equilibrium model, by considering users' behaviour and operators' costs, and describing how supply and demand interact under different market conditions. The previous chapter has highlighted the double externality effect that a supplementary frequency can generate in railways: "expected schedule delay effect" (Mohring effect) or a "congestion effect".

Traditionally, and as described in chapter 2, transport economic research has put great effort on studying the negative externality, or the "congestion effect". The possibility of using pricing to deal with congestion externalities, as in the road sector, has also been developed for modes of transport programmed in advance, particularly in the air sector.

A major difference between congestion on roads and in aviation is that, typically, individual road users do not have market power. In contrast, in aviation, recent airport research has explicitly recognised that airlines operate under imperfect conditions of competition and these characteristics must be considered in the optimal pricing recommendations. For example, if their market share is high, airlines will internalise part of their own congestion externalities that must be taken into account in the implementation of congestion pricing (Daniel, 1995; Brueckner, 2002, 2005).

Nevertheless, empirical evidence of the self-internalisation hypothesis is still subject to debate (Mayer and Sinai, 2003; Morrison and Winston, 2007). Consequently, one of the most controversial issues in airport congestion pricing in recent decades has been that of determining whether an airline structure can be treated as atomistic in the theoretical formalisation or whether self-imposed congestion must be considered. As described in the interpretive review of Zhang and Czerny (2012), recent airport research has focused on analysing how airport economics and policy recommendations should incorporate strategic interactions between airlines with market power.

In this context, this market power effect has been investigated, for example, by Pels and Verhoef (2004), who developed an airport pricing model considering the specificities of air transport (market power, partial congestion internalisation and multiple regulatory authorities), in order to establish if congestion pricing is a useful policy tool under these conditions. Villemeur et al. (2015) used a different perspective to present a realistic model to seek an optimal buffer-time. They used a methodology to estimate the social cost of delays implementing a simple calibration model.

In contrast to this rich debate on the airline sector, focusing on the factors which would determine how pricing can be an optimal tool for solving congestion, very few academic papers have considered congestion in rail transport. Some notable exceptions are the High Level Group on infrastructure charging (Nash and Samson, 1999), papers by Quinet (2003) and Nash and Matthews (2003) and the extended analysis performed in Great Britain to define a rail capacity charge (ARUP & Network Rail, 2013; Haith, 2015). The latter specifies the case of pricing railway congestion from a theoretical viewpoint.

The aim of this chapter is to analyse the interactions between demand and supply, as it has been done for other transport modes, and discuss if standard theoretical conclusions in the short-run, such as price mechanisms, are optimal tools for dealing with rail congestion externalities. Moreover, it examines under which conditions these tools are optimal, taking into account the specificities of the rail sector, developed in the previous chapters.

4.2 Notation and Assumptions

For the general specification of the model, a number of assumptions are made that we present below.

Assumption 1: Demand function

The inverse aggregate demand is linear in form:

$$D(N) = A + BN \tag{4.1}$$

where $A > 0$ and $B < 0$. A represents the maximum reservation price for the rail route and B is the demand sensitive parameter. $D(N)$ represents the marginal passenger's maximum willingness to pay for the rail service, including monetised time costs (with $N > 0$ and $D(N) > 0$)

Assumption 2: Frequency

We consider that the train operating company's (TOC) frequency is given by the passenger demand:

$$f = \frac{N}{k} \tag{4.2}$$

where k is the average number of passengers per train (the product of the load factor and seat capacity are considered as given), we can rewrite the result equation in chapter 3 as:

$$GC_0\left(\frac{N}{k}\right) \quad = \quad \alpha T \quad + \quad \frac{k}{N}\frac{\beta\gamma}{2(\gamma+\beta)} \quad + \quad \delta\left[\frac{\mu_{d_{1,i}}^2+\sigma_{d_{1,i}}^2}{2(\frac{k}{N}-\frac{1}{f_{max}})}+\mu_{d_{1,i}}\right] \tag{4.3}$$

where:

- αT is planned travel time cost.
- $\frac{k}{N}\frac{\beta\gamma}{2(\gamma+\beta)}$ is the expected schedule delay cost (when there is no travel time variability).
- $\delta\left[\frac{\mu_{d_{1,i}}^2+\sigma_{d_{1,i}}^2}{2(\frac{k}{N}-\frac{1}{f_{max}})}+\mu_{d_{1,i}}\right]$ is the random delay cost.

Assumption 3: Generalised cost for users

The generalised cost of a train's service as experienced by passengers is characterised by a generalised user cost function $GC(p, GC_0)$ where p is the fare and GC_0 is the monetised time cost function for the users. The generalised user cost function is linearly additive in form:

$$GC = p + GC_0(N/k) \tag{4.4}$$

Assumption 4. Supply function: Trains Operator Company (TOC).

The cost of the TOC (subscript O denotes the operator) is composed of an operating cost per train c_O^f considered constant, a congestion cost c_O^d depending on the average delay, a fixed cost F_O and a toll per train τ_O^f. The total operating cost for the TOC is therefore:

$$C_O(f) = c_O^f f + c_O^d f T_d(f) + F_O + \tau_O f \tag{4.5}$$

which may be rewritten as

$$C_O\left(\frac{N}{k}\right) = \frac{N}{k}\left(c_O^f + c_O^d T_d\left(\frac{N}{k}\right) + \tau_O\right) + F_O \tag{4.6}$$

where $T_d(f)$ is the random delay per train and which may be rewritten as $T_d(f) = \left[\frac{\mu_{d_{1,i}}^2+\sigma_{d_{1,i}}^2}{2(\frac{N}{k}-\frac{1}{f_{max}})}+\mu_{d_{1,i}}\right]$

4.3 Demand and Supply Equilibriums

Considering these assumptions it is now possible to calculate several analytical and graphical results. We consider that there are three types of actor: passengers, the train operating companies (TOC) and a regulatory authority, each having a specific maximisation problem.

The model is solved in three steps. Firstly, the passenger demand function for train services is defined. Then, considering this demand function, the TOC problem is described and the associated profit maximisation optimality conditions are derived. Finally, the regulator's maximisation problem is described.

4.3.1 The monopoly equilibrium: analytical solution

In this section, we focus on a simplified network with a double track line with homogeneous traffic between two train stations. The train service is provided by a single integrated TOC in a monopoly market situation. Although this hypothesis may seem restrictive, it represents the reality of certain rail transport services.

Traditionally, and until the end of the 20th century, the rail sector was considered a natural monopoly. Even today, a single operator supplies national high speed and regional and local services with a dedicated infrastructure in numerous European countries.

Further research could lead to relaxing some restrictions and consider the possibility of multi-modal competition for example, or the separation between the infrastructure manager (IM) and the TOC, as has been the case in many European countries since the end of the 1990s. [1]

The passenger optimisation problem

The marginal passenger's maximum willingness to pay for the rail service, including monetised time costs, is given by equation 4.1, while the user's generalised cost of travelling is given by equation 4.4. Considering Wardrop's equilibrium conditions, marginal benefits are equal to the generalised cost at equilibrium.

Formally, at equilibrium

[1] "Member States shall take the measures necessary to ensure that the accounts for business relating to the provision of transport services and those for business relating to the management of railway infrastructure are kept separate." (Directive 91/440/EEC, Article 6,)

$$p + GC_0\left(\frac{N}{k}\right) = D(N) \tag{4.7}$$

The passenger equilibrium condition in this case implies the following fare:

$$p = A + BN - GC_0\left(\frac{N}{k}\right) = A + BN - \alpha T - \frac{k}{N}\frac{\beta\gamma}{2(\gamma+\beta)} - \delta\left[\frac{\mu_{d_{1,i}}^2 + \sigma_{d_{1,i}}^2}{2(\frac{k}{N} - \frac{1}{f_{max}})} + \mu_{d_{1,i}}\right] \tag{4.8}$$

The TOC maximisation problem

When there is only one firm on the market, it is very unlikely to take the market price as given. Instead, the monopoly recognises its influence over the market price and determines the level of price and output that maximises its profits.

The integrated operator (infrastructure manager+ train operator company) maximises its profit with respect to N, i.e. the number of passengers.

The maximisation problem for the TOC company is:

$$\max_N \Pi = \left(A + BN - GC_0\left(\frac{N}{k}\right)\right)N - \frac{N}{k}\left(c_O^f + c_O^d T_d\left(\frac{N}{k}\right) + \tau_O\right) - F_O \tag{4.9}$$

The necessary first order conditions are:

$$A + BN - GC_0(\frac{N}{k}) + N\left(B - \frac{\partial GC_0(\frac{N}{k})}{\partial N}\right) - \frac{c_O^f}{k} - \frac{\tau_O}{k} - \frac{c_O^d\, T_d(\frac{N}{k})}{k} - \frac{c_O^d\, N\left(\frac{\partial T_d(\frac{N}{k})}{\partial N}\right)}{k} = 0 \tag{4.10}$$

Each additional passenger transported by the TOC generates a marginal cost for the operator corresponding to $\frac{c_O^f}{k} + \frac{\tau_O}{k} + \frac{c_O^d\, T_d(\frac{N}{k})}{k} + \frac{c_O^d\, N\left(\frac{\partial T_d(\frac{N}{k})}{\partial N}\right)}{k}$, where the first three terms represent the marginal operating cost per passenger and the fourth term represents the marginal direct congestion cost, namely the increase in operating costs for a supplementary passenger transported. In addition, the TOC is subject to a change in its income, dependent on the change in the generalised user cost term $\frac{\partial GC_0(N)}{\partial N}$. This term can be positive or negative, depending whether it represents an expected schedule effect or a congestion effect (Chapter 3). It can the stand for an indirect cost or benefit for the firm, reducing or increasing the passenger's willingness to pay.

We can derive the equilibrium fare from the first order condition and the passenger equilibrium solution:

$$p^M = \frac{1}{k}\left[c_O^f + \tau_O + c_O^d\, T_d\left(\frac{N}{k}\right)\right] + \frac{c_O^d\, N\, \left(\frac{\partial T_d(\frac{N}{k})}{\partial N}\right)}{k} + N\frac{\partial GC_0(\frac{N}{k})}{\partial N} - NB \tag{4.11}$$

where the first term in brackets represents the TOC's cost per passenger, the second and third terms reflect the firm's-internal direct and indirect congestion costs respectively and the last term represents the traditional monopoly mark-up ($B < 0$).

From the first order condition for profit maximisation, it is clear that the monopoly completely internalises the costs incurred by itself and its passengers due to an additional traveller. Otherwise, the equilibrium fare equation shows monopoly prices over marginal costs using a mark-up.

Monopolistic behaviour is not Pareto efficient, not because of externalities (which are fully internalised) but because its mark-up is linked to its market power.

According to basic economic theory, competitive industry operates at an optimal point N_{PMC} where price equals marginal cost. In our case, the integrated monopoly chooses an output N_M lower than the competitive output N_{PMC}, and with a higher price. For this reason, consumers will typically be worse off in an industry organised as a monopoly than in one organised competitively.

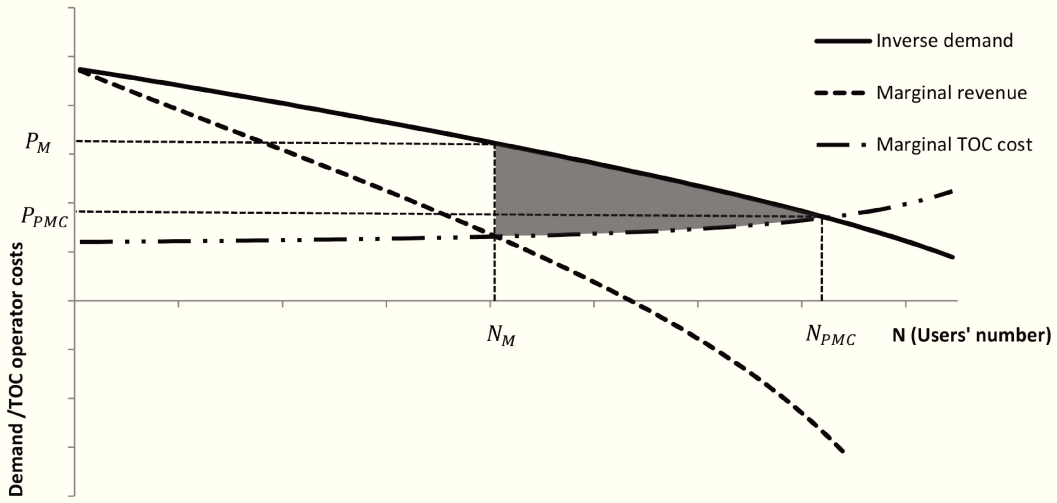

Figure 4.1: Inefficiency of monopoly

Figure 4.1 illustrates the deadweight loss due to monopoly (shadow area), estimating the value of one unit of lost output at the price that people are willing to pay for that unit.

The regulator's maximisation problem

Without budget constraint

In the previous section we studied the monopoly maximisation problem in a non-regulated scenario. As explained, monopoly behaviour is not Pareto efficient. In this section we consider a welfare-maximising regulation.

The regulator's objective function is to maximise the social surplus for the entire network: the regulator considers consumer surplus and monopoly profit. The regulator's maximisation problem is written as:

$$\max_N W = \int_0^N (A + Bn)dn - cg_0\left(\frac{N}{k}\right)N - \frac{N}{k}\left(c_O^f + c_O^d T_d\left(\frac{N}{k}\right)\right) - F_O \tag{4.12}$$

The first order condition yields:

$$A + BN - GC_0\left(\frac{N}{k}\right) - N\left(\frac{\partial GC_0(\frac{N}{k})}{\partial N}\right) - \frac{c_O^f}{k} - \frac{c_O^d\, T_d(\frac{N}{k})}{k} - \frac{c_O^d\, N\, \left(\frac{\partial T_d(\frac{N}{k})}{\partial N}\right)}{k} = 0 \tag{4.13}$$

We can derive the equilibrium fare from the first order conditions and the passenger equilibrium solution:

$$p^W = \frac{1}{k}\left[c_O^f + c_O^d\, T_d\left(\frac{N}{k}\right)\right] + \frac{c_O^d\, N\, \left(\frac{\partial T_d(\frac{N}{k})}{\partial N}\right)}{k} + N\frac{\partial GC_0(\frac{N}{k})}{\partial N} \tag{4.14}$$

The regulator would set a price equal to the marginal costs (train company operating cost per passenger and its internal direct and indirect congestion costs).

By comparing the first order conditions for profit maximisation and welfare maximisation we observe:

$$\frac{\partial W}{\partial N} - \frac{\partial \Pi}{\partial N} = -BN + \frac{\tau_O}{k} \tag{4.15}$$

To calculate the welfare optimising toll per train, equation 4.15 must equal zero.

$$\tau^W = BNk \tag{4.16}$$

As $B < 0$, equation 4.16 should be interpreted as an optimal subsidy per user. The optimal subsidy would incite the monopoly to provide the rail service at the social optimal level, but also lead to overdraft financing. This solution can be considered as a first-best analysis.

Until now, the regulator maximisation problem has been considered without constraints and/or market distortions. In the current financial and budgetary context, it seems reasonable to consider that public subsidies are not unlimited. Consequently, and in order to propose realistic policy insights, a second-best analysis must also be explored.

With budget constraint

In fact, public subsidies are not unlimited and it is costly to raise public funds because taxes are distortionary. This additional cost is known as the opportunity cost of public funds (OCPF), defined as the cost to a society of raising one euro of tax revenue. Boiteux (1956) raised the issue of modifying pricing which maximises the welfare function and ensures also the budgetary equilibrium of the firm. The regulator maximisation problem is the same in equation 4.12 but takes the TOC's budgetary constraint into account:

$$\left(A + BN - GC_0\left(\frac{N}{k}\right)\right)N - \frac{N}{k}\left(c_O^f + c_O^d T_d\left(\frac{N}{k}\right) + \tau_O\right) - F_O \geq 0 \tag{4.17}$$

In order to solve this, we can write the Lagrangian equation:

$$\begin{aligned}\mathcal{L} = &\int_0^N (A + Bn)dn - cg_0\left(\frac{N}{k}\right)N - \frac{N}{k}\left(c_O^f + c_O^d T_d\left(\frac{N}{k}\right)\right) - F_O \\ &+ \lambda\left[\left(A + BN - GC_0\left(\frac{N}{k}\right)\right)N - \frac{N}{k}\left(c_O^f + c_O^d T_d\left(\frac{N}{k}\right) + \tau_O\right) - F_O\right]\end{aligned} \tag{4.18}$$

with λ being the Lagrange multiplier of the budgetary constraint, indicating by how much the social profit would increase if the desired profit or authorised deficit for the TOC were decreased by a unit, or in other words, it is a parameter reflecting the opportunity cost of public funds. The opportunity cost of public funds is considered higher than one (being the budget constraint parameter $\lambda < 1$).

By solving the optimal price under budget constraints, and comparing it to the equation 4.14, we obtain:

$$p^W_{constraint} - p^W = -\frac{\lambda BN}{\lambda + 1} \tag{4.19}$$

which is a positive term, knowing that the inverse demand sensitive parameter is negative ($B < 0$). Indeed, this means that the optimal price under budgetary constraint is higher than in the first case. When the opportunity cost of public funds is not zero, the regulator will propose higher prices for users, meaning that the share between the users and taxpayers for covering TOC costs will be different.

It is also interesting to note that the higher B is, the higher the difference between both prices under different conditions would be. This means that the lower the demand sensitive parameter is, the higher the optimal price under budget constraints will be. Furthermore, we observe that the difference between both optimal prices also depends positively on λ. If the opportunity cost of public funds increases, the difference between optimal price with and without budget constraint will also increase.

Comparing the FOC for profit maximisation and welfare maximisation under budget constraints, we obtain an optimal subsidy under constraint:

$$\tau^W_{constraint} = \frac{kBN}{\lambda + 1} \tag{4.20}$$

Two important conclusions can be drawn when comparing both optimal tolls with or without budgetary constraint. Firstly, the optimal toll in a first-best world (equation 4.16) is higher than in a second-best scenario considering budget distortions (equation 4.20). This result is intuitive, in a second-best world the regulator will subsidise at a lower level if public funds are costly. Secondly, the higher the opportunity cost of public funds is (if λ increases), the lower the optimal subsidy will be.

If the opportunity cost of public funds is equal to zero, there will be no difference in the subsidy in the two situations.

4.3.2 The symmetric duopoly equilibrium: analytical solution

In this section, we consider that the train service is provided by two symmetric firms in the market and that they produce a homogeneous product (in our particular case, the same train service between two cities).

We assume that the two firms are simultaneously trying to decide what quantity to produce, taking the other train company's output as given. They act as Cournot duopolists.

The passenger optimisation problem

As in the previous section, considering Wardrop's equilibrium conditions, the marginal benefits are equal to the generalised cost at equilibrium.

Formally, at equilibrium

$$p + GC_0\left(\frac{N}{k}\right) = D(N) \tag{4.21}$$

The passenger equilibrium condition implies the following fares for operator i $(i = 1, 2)$:

$$p_i = A + B(N_1 + N_2) - \alpha T - \frac{k}{(N_1 + N_2)}\frac{\beta\gamma}{2(\gamma+\beta)} - \delta\left[\frac{\mu_{d_{1,i}}^2 + \sigma_{d_{1,i}}^2}{2(\frac{k}{(N_1+N_2)} - \frac{1}{f_{max}})} + \mu_{d_{1,i}}\right] \tag{4.22}$$

The TOC maximisation problem

In this section operators' maximise their profits with respect to N_i, taking the competitor quantity as given.

Each TOC maximises profit with respect to N_i, i.e. its number of passengers.

The maximisation problem for the TOC company $i(i = 1, 2)$ is:

$$\begin{aligned}\max_{N_i}\Pi = &\left[A + B(N_1 + N_2) - GC_0\left(\frac{(N_1 + N_2)}{k}\right)\right]N_i \\ &- \frac{N_i}{k}\left(c_O^f + c_O^d T_d\left(\frac{(N_1 + N_2)}{k}\right) + \tau_O\right) - F_O\end{aligned} \tag{4.23}$$

The necessary first order conditions are:

$$\begin{aligned}&A + B(N_1 + N_2) - GC_0\left(\frac{(N_1 + N_2)}{k}\right) + N_i\left(B - \frac{\partial GC_0(\frac{N1+N2}{k})}{\partial N_i}\right) \\ &- \frac{c_O^f}{k} - \frac{\tau_O}{k} - \frac{c_O^d\, T_d(\frac{(N_1+N_2)}{k})}{k} - \frac{c_O^d\, N_i\left(\frac{\partial T_d(\frac{(N_1+N_2)}{k})}{\partial N_i}\right)}{k} = 0\end{aligned} \tag{4.24}$$

Solving the FOC yields the following equilibrium output:

$$N_1 = N_2 = \frac{\tau_O + c_O^d\, T_d(\frac{(N_1+N_2)}{k}) + k\, GC_0\left(\frac{(N_1+N_2)}{k}\right) + c_O^f - k\, A}{k\left(3\, B - \frac{\partial GC_0(\frac{N1+N2}{k})}{\partial N_i}\right) - c_O^d\left(\frac{\partial T_d(\frac{(N_1+N_2)}{k})}{\partial N_i}\right)} \tag{4.25}$$

We can derive the equilibrium fare from the first order condition and the passenger equilibrium solution, :

$$p_i^D = \frac{1}{k}\left[c_O^f + \tau_O + c_O^d\, T_d\left(\frac{(N_1+N_2)}{k}\right)\right] + \frac{c_O^d\, N_i\left(\frac{\partial T_d(\frac{(N_1+N_2)}{k})}{\partial N_i}\right)}{k} + N_i \frac{\partial GC_0(\frac{(N_1+N_2)}{k})}{\partial N_i} - N_i B \tag{4.26}$$

where the first term in brackets represents the TOC's cost per passenger as in the previous case, the second and third terms reflect the firm's-internal direct and indirect congestion costs, respectively, and the last term represents the market power effect. Unlike the previous section, the first order condition shows clearly that the train company i internalises only the congestion incurred by itself and its passengers. In our symmetric duopoly equilibrium case, we can say that train operator companies internalise only half of the congestion they cause.

The regulator maximisation problem

The duopoly maximisation problem shows that TOCs set their prices considering their marginal cost plus their market mark-up, but they do not internalise all the externalities associated with an additional passenger. Ignoring the externalities imposed on the other TOCs and the market power effect are not consistent with efficient pricing. Regulatory strategies must be considered to deal with these inefficiencies.

The regulator objective function is to maximise the social surplus for the entire network: the regulator considers the consumer surplus and profit of both operators. The regulator maximisation problem is:

$$\begin{aligned}\max_{N_i} W = \int_0^{N1+N2} (A + Bn)dn - cg_0(\frac{(N1+N2)}{k})(N1+N2) \\ - \frac{(N1+N2)}{k}(c_O^f + c_O^d T_d(\frac{(N1+N2)}{k})) + F_O\end{aligned} \tag{4.27}$$

The first order condition yields:

$$A + BN - GC_0(\frac{(N1+N2)}{k}) - (N1+N2)\left(\frac{\partial GC_0(\frac{(N1+N2)}{k})}{\partial N_i}\right) - \frac{c_O^f}{k} - \frac{c_O^d\, T_d(\frac{(N1+N2)}{k})}{k} - \frac{c_O^d\, N\left(\frac{\partial T_d(\frac{(N1+N2)}{k})}{\partial N_i}\right)}{k} = 0 \quad (4.28)$$

We can derive the equilibrium fare from the first order condition and the passenger equilibrium solution:

$$p_i^{W2} = \frac{1}{k}\left[c_O^f + c_O^d\, T_d(\frac{(N1+N2)}{k})\right] + \frac{c_O^d\,(N1+N2)\left(\frac{\partial T_d(\frac{(N1+N2)}{k})}{\partial N_i}\right)}{k} + (N1+N2)\frac{\partial GC_0(\frac{(N1+N2)}{k})}{\partial N_i} \quad (4.29)$$

The regulator sets a price equal to the marginal costs (train company operating cost per passenger and both firms'-internal direct and indirect congestion costs).

By comparing the first order conditions for welfare maximisation and profit maximization we observe:

$$\frac{\partial W}{\partial N_i} - \frac{\partial \Pi}{\partial N_i} = \frac{\tau_O}{k} - \frac{c_O^d\, N_{-i}\left(\frac{\partial T_d(\frac{(N1+N2)}{k})}{\partial N_i}\right)}{k} - N_{-i}\frac{\partial GC_0(\frac{(N1+N2)}{k})}{\partial N_i} - BN_i \quad (4.30)$$

In a symmetric equilibrium $N_{-i} = N_i$

The welfare optimising toll must be calculated so that equation 4.30 equals zero.

$$\tau_0^W = N_{-i}\, c_O^d\left(\frac{\partial T_d(\frac{(N1+N2)}{k})}{\partial N_i}\right) + kN_{-i}\frac{\partial GC_0(\frac{(N1+N2)}{k})}{\partial N_i} + kBN_i \quad (4.31)$$

where $N_{-i}\, c_O^d\left(\frac{\partial T_d(\frac{(N1+N2)}{k})}{\partial N_i}\right)$ represents the direct marginal cost for TOC_{-i} not internalised by TOC_i, and $N_{-i}\frac{\partial GC_0(\frac{(N1+N2)}{k})}{\partial N_i}$ represents the indirect externality on passengers for TOC_{-i} not internalised by TOC_i. This externality can be positive or negative, as exposed in chapter 3, representing an indirect cost or a benefit for TOC_{-i}, reducing or incrementing the passenger's willingness to pay. BN_i represents the market power for

TOC_i.

In fact, the toll sign (subsidies or additional pricing) will depend on the combination of three parameters presented above (equation 4.31). For example, on a very congested line, with $\frac{\partial GC_0(\frac{(N1+N2)}{k})}{\partial N_i} > 0$, the regulator should apply a positive toll if the sum of both external costs (direct for the TOCs and indirect for the passengers) is higher than the market power effect.

Substituting the toll rule 4.31 in the optimal output 4.25 and considering that $N_{-i} = N_i$ in the symmetric equilibrium, yields the optimal quantity per TOC :

$$N_i^w = \left[\frac{+c_O^d \, T_d(\frac{(N_1+N_2)}{k}) + k \, GC_0 \left(\frac{(N_1+N_2)}{k}\right) + c_O^f - k \, A}{2 \left(k \, B - k \, \frac{\partial GC_0(\frac{N1+N2}{k})}{\partial N_i} - c_O^d \, \frac{\partial T_d(\frac{(N_1+N_2)}{k})}{\partial N_i} \right)} \right] \tag{4.32}$$

4.3.3 Generalised case equilibrium: analytical equilibrium solution

Now let us assume that we have several firms involved in a Cournot equilibrium, not just two. As previously, we consider that each firm considers the other firms output choices as given and they maximise their profit.

We assume that there are M firms and that $N = N_1 + N_2 + ...N_M$. Market shares of firms are symmetric and depend on the total number of firms $m = \frac{1}{M}$, so $N_i = Nm$

Considering the results in the previous section, we can generalise the conclusions on the duopoly case to a more extended application.

$$\tau_0^W = \frac{N(1-m) \, c_O^d \left(\frac{\partial T_d(\frac{(N)}{k})}{\partial N_i}\right)}{k} + N(1-m)\frac{\partial GC_0(\frac{(N)}{k})}{\partial N_i} + BNm \tag{4.33}$$

From this generalised expression we can observe that the higher the number of train companies is, the lower their market effect will be. Furthermore, the larger the number of companies is, the larger the externality non internalised by each TOC. As exposed in the first section, if there is a monopoly ($M = 1$) , the first two terms are equal to zero, and the externalities are fully internalised.

To conclude, the choice of the level of the toll will depend on the market composition (number of firms) and on the service line typology: does an additional passenger generate an expected schedule delay externality effect or a congestion externality effect?

4.4 A Few Graphical Illustrations

The previous section defined the supply-demand equilibriums analytically, by considering different market scenarios. In order to further illustrate the properties of the model, this section will present several graphical results.

It is not the purpose of this section to precisely describe a real-life rail system. The parameters therefore do not need to correspond to real life values.

Figure 4.2 compares the total surplus for the three main market scenarios: monopoly and duopoly profit maximisation and regulator welfare maximisation equilibrium, by considering different demand sensitive parameters. We observe that the total surplus in the regulatory situation is always higher than the two others, independently of the inverse demand sensitive parameter.

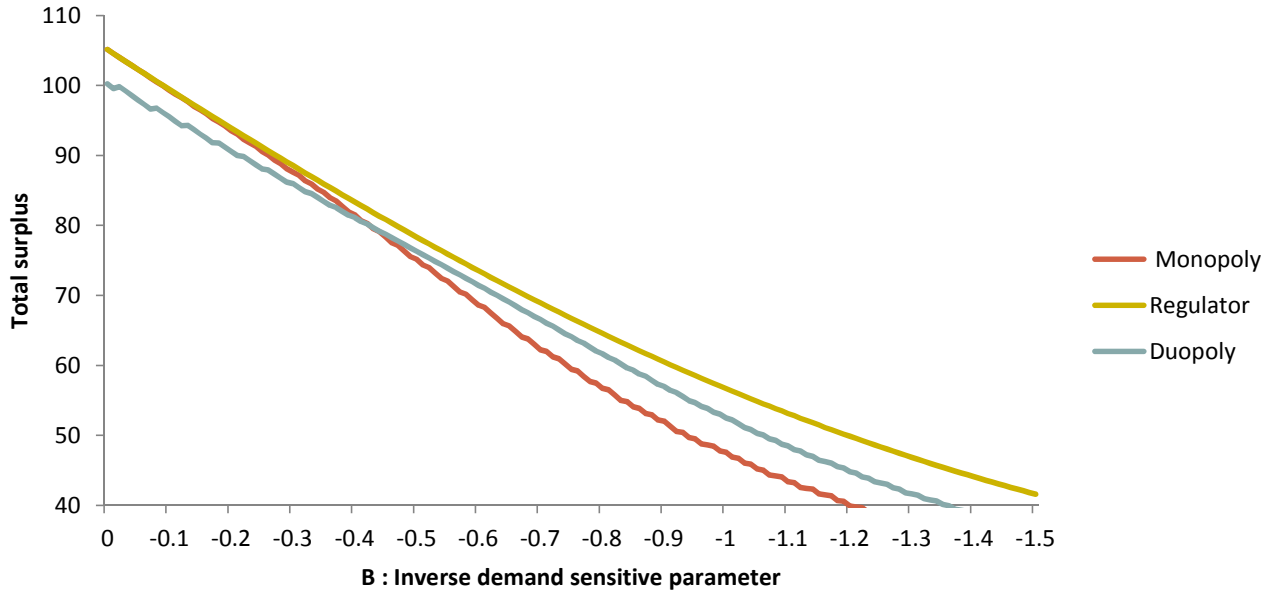

Figure 4.2: Impact of the inverse demand sensitive parameter on total surplus.

Nevertheless, the level of surplus between the monopoly and the duopoly situation depends on the inverse demand sensitive parameter. As stated previously, a monopoly completely internalises the externalities but it has the power to impose a mark-up. Alternatively, the duopoly presents a lower mark-up (increasing total surplus) but the two firms do not internalise externalities completely (diminishing total surplus) and we have to consider double fixed costs from a global perspective (diminishing total surplus). The

total surplus "battle" between the monopoly and the duopoly market situation will depend on the impact of fixed costs and externalities versus the impact of market power.

Indeed, when the market power is low (low B), the monopoly does not have sufficient power to discriminate and the consumer surplus is similar to a competition situation (we can see in the figure that initially, the monopoly and regulator curves are symmetric). However, in the duopoly situation, we have to consider the double fixed costs in the total surplus calculation. At the start of the curves, the higher operating costs of having two train companies is not offset by the extra consumer surplus of having some kind of competition.

On the other hand, when the market power is high (high B) and the capacity of the monopoly to discriminate is strong, the presence of a second train company increases the consumer surplus. In this part of the curve, the double operating costs of the duopoly are offset by the gain in consumer surplus, and the total duopoly surplus is higher than that of the monopoly.

The previous arguments beg the question of natural monopoly and fixed costs. In some situations, when there are high fixed costs and low marginal costs (this kind of situation often arises with public facilities) a single operating company is more efficient than having several. Consequently, Figure 4.3 compares the total surplus for the three previous market situations depending on fixed costs. As predicted, when fixed costs are high, the monopoly surplus is higher than the duopoly surplus.

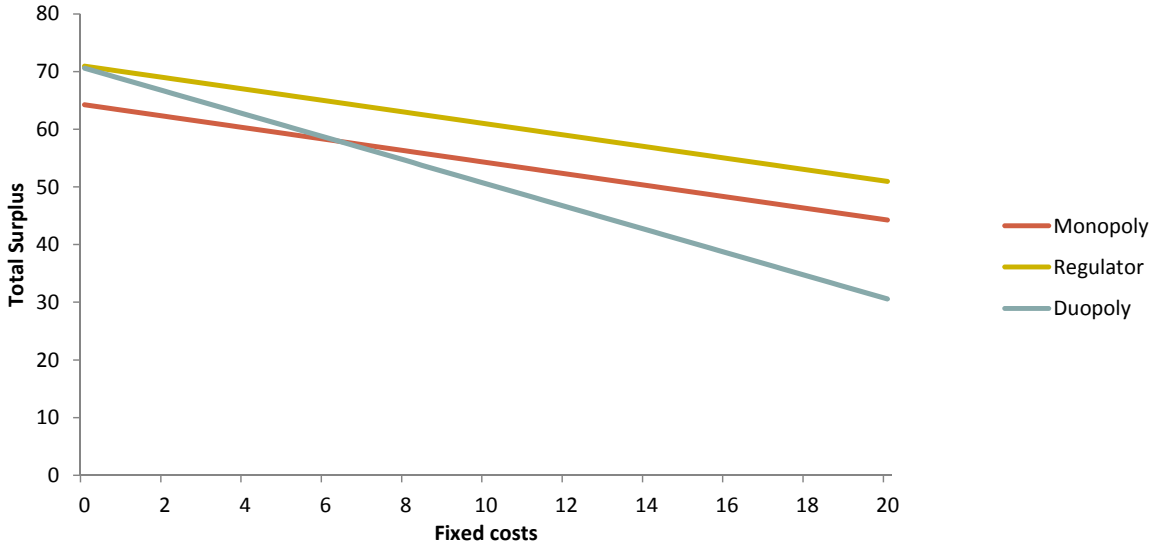

Figure 4.3: Impact of fixed costs on the total surplus.

To conclude, it is important to bear in mind that the total surplus hierarchy between a monopoly and a duopoly depends on the market power and on the fixed costs. The trade-off between these two situations will depend on the gain for consumers of having a second operating company (depending on the market power) and the costs of doubling the initial investment on fixed costs.

In practice, in a competitive inter-modal market, where the railway market power is low, it might be beneficial for the community that only one railway company operates on the network. Nevertheless, this conclusion depends also on the level of fixed costs. If a new operator could enter the market without bearing high fixed costs (we can imagine for example by renting or leasing its rolling stock), the conclusion would be different, pleading for several TOCs in the market.

As exposed in the previous section, an optimal toll should be determined in some situations in order to adjust the inefficiencies associated with market power or the non-considered externalities. Figure 4.4 compares the impact of inverse demand sensitive parameter on the three toll scenarios: monopoly (considering a budget constraint or not) and duopoly.

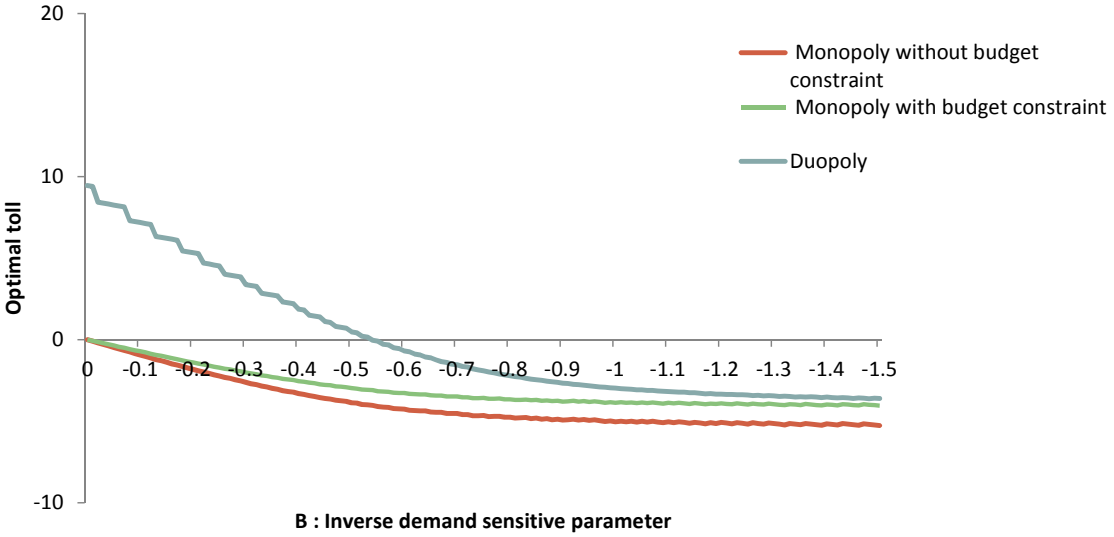

Figure 4.4: Impact of inverse demand sensitive parameter on the optimal toll.

Firstly, we observe that the higher the inverse demand sensitive parameter is, the higher a firm's market power; thus the trend is towards a negative toll (therefore a subsidy). In the case of the monopoly situation, the optimal toll is always negative: as a monopoly fully internalises congestion, the higher its market power and thus its ability

to discriminate, and the higher the optimal subsidy which will incite the monopoly to provide the rail service at the socially optimal level.

In the case of a duopoly, the sign of the optimal toll strongly depends on the inverse demand sensitive parameter. As shown in the analytical results in equation 4.31, the toll sign (subsidy or additional pricing) will depend on the combination of the three parameters described. If the line presents a high inverse demand sensitive parameter, the negative market power term will be larger than the congestion terms (direct and indirect).

From the practical viewpoint, when firms present strong market power its seems more optimal to subsidise activity than adding an extra tax, in order to avoid worsening the initial distortion associated with the traditional monopoly's mark-up.

Furthermore, it is interesting to note that in a second best world, considering budget constraints, the optimal subsidy is always lower than in the monopoly's first best world. If subsidies are costly, the regulator will subsidise at a lower level. Figure 4.5 reflects this situation in detail, by comparing optimal subsidies with a variation of the Lagrange multiplier of the budgetary constraint. For a given market power situation, the higher the opportunity costs of public funds are, the lower the subsidy will be: when public funds are costly, the share between consumers and taxpayers is modified, because a trade-off exists between consumer surplus and scarce public funds.

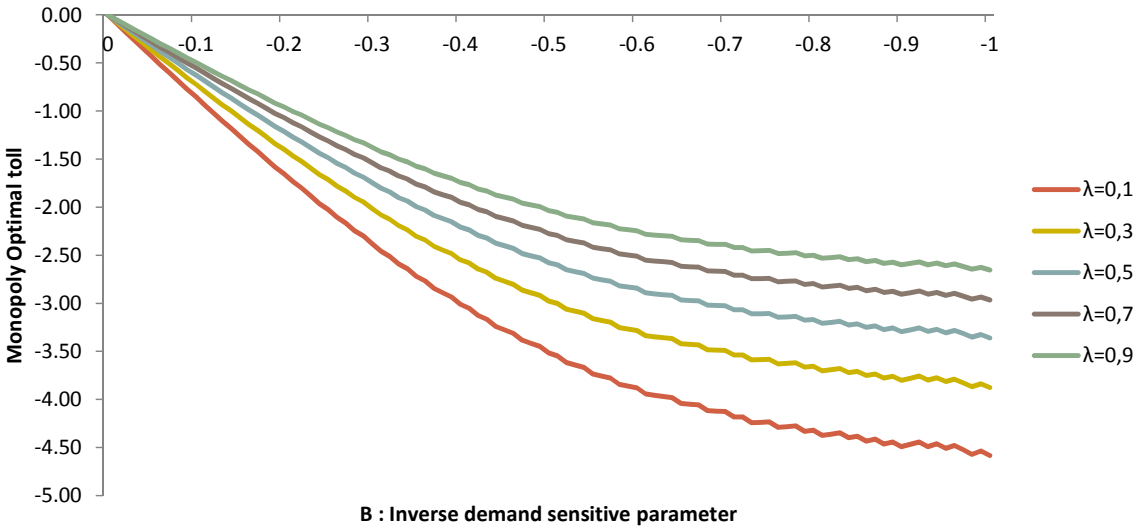

Figure 4.5: A change in the Lagrangian multiplier: the impact of the inverse demand sensitive parameter on the monopoly optimal toll.

As exposed previously, the sign and the level of the optimal toll in a duopoly situation will depend on the combination of three parameters: direct congestion costs, indirect congestion costs and market power mark-up. Figure 4.6 shows the evolution of these three terms as a function of the initial average delay. All other things being equal, when the initial average delay is low, congestion terms are low, so the negative sign of the total toll is given by the market power effect. In contrast, when the initial average delay is high, the congestion terms, and especially indirect congestion, are the main components of the toll sign and amount.

If the quality of service of a line is deteriorated (high average initial delay), it seems important for the regulator to impose a positive toll. This optimal toll will allow the TOC to internalise the indirect congestion costs supported by the users of the competing TOC. In a poor reliability scenario, the market power effect on the toll is insignificant compared to the externality effect.

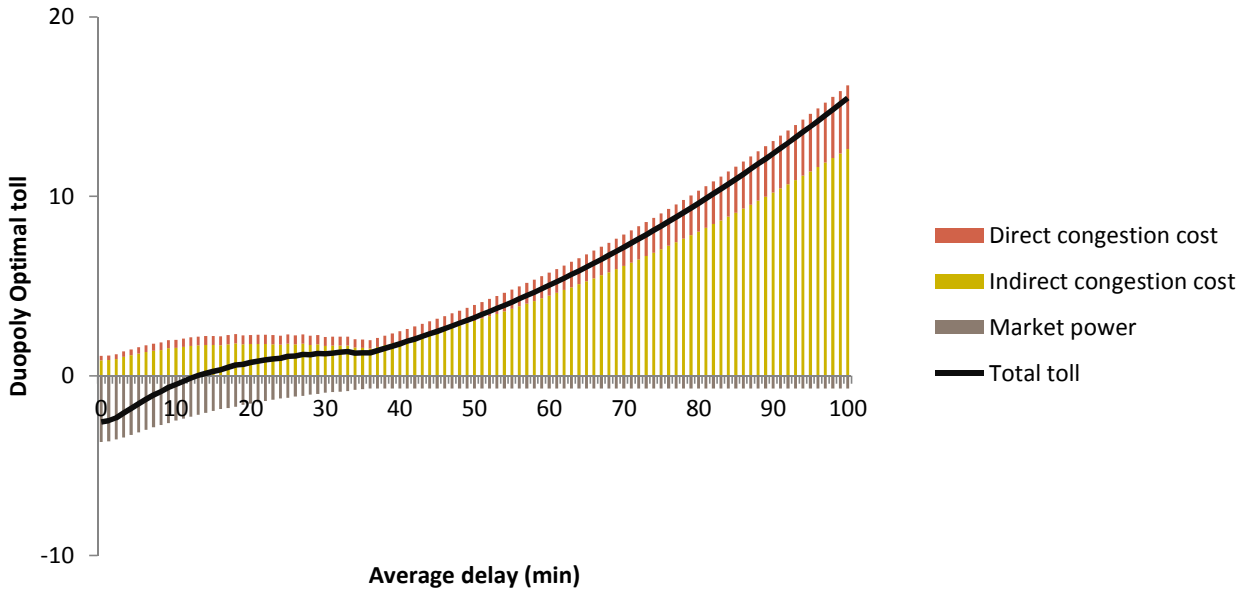

Figure 4.6: Impact of average delay on the duopoly optimal toll.

It is interesting to compare the combination of the three terms as a function of frequency (or the number of users) in Figure 4.7. Considering the other parameters as given, we observe that the optimal toll is negative in a low frequency scenario. In fact, when frequencies are low, and as exposed in chapter 3, an additional train decreases the schedule delay effect (called "Mohring effect), generating a positive externality. In this context, the regulator should subsidise the rail service.

Otherwise, in dense areas where we expect high frequencies, an additional train would be translated into an additional congestion effect for users and other TOCs. In this case, the regulatory authority should impose a positive congestion toll (corrected by market power) in order to incite the operators to internalise the costs generated for other train companies.

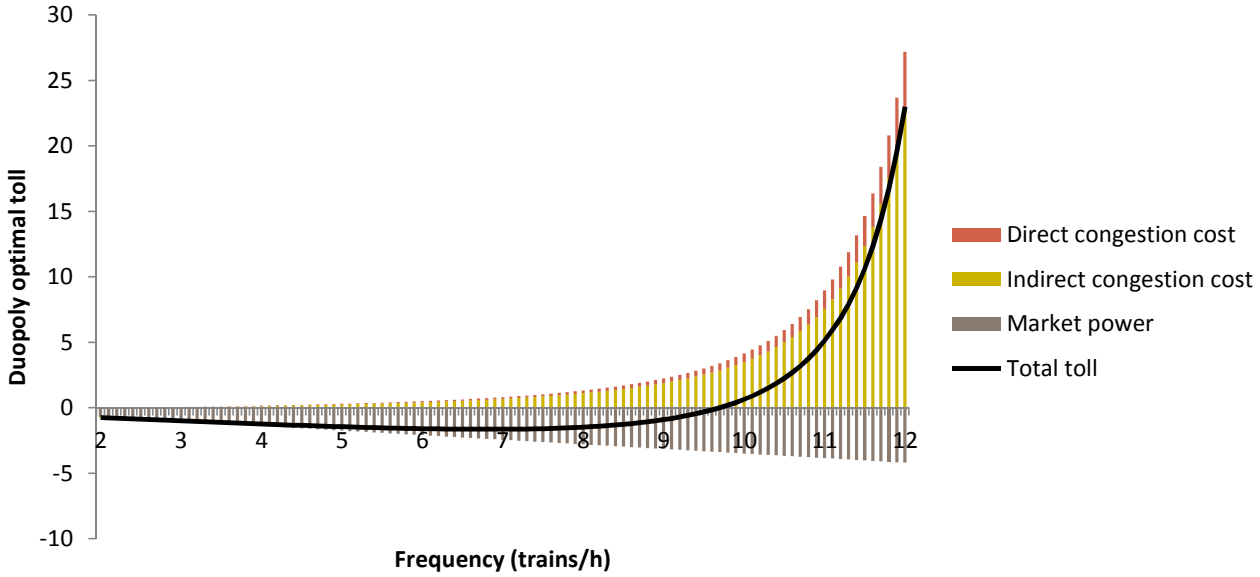

Figure 4.7: Impact of frequency on the duopoly optimal toll.

To complete the analysis, it is interesting to analyse the relationship between the optimal toll and frequency described in figure 4.8, considering a variation in the average initial delay and inverse demand sensitive parameter.

This figure shows that in lines with reliability problems (high average initial delays), the optimal positive toll starts at lower frequency levels. If we consider a line with a poor level of reliability, an additional train will generate congestion costs at a lower frequency level than for a good quality line. Indeed, the level of high frequency that justifies a positive congestion toll will also depend on the initial quality of the line. The same frequency on two different lines does not generate the same congestion costs; it depends on the importance of the initial average delay.

Furthermore, the sign on the optimal toll for a given frequency will also depend on market power (Figure 4.9). On lines with strong inter-modal competition, where the rail market power is low, the optimal toll will be predominantly positive (except in situations

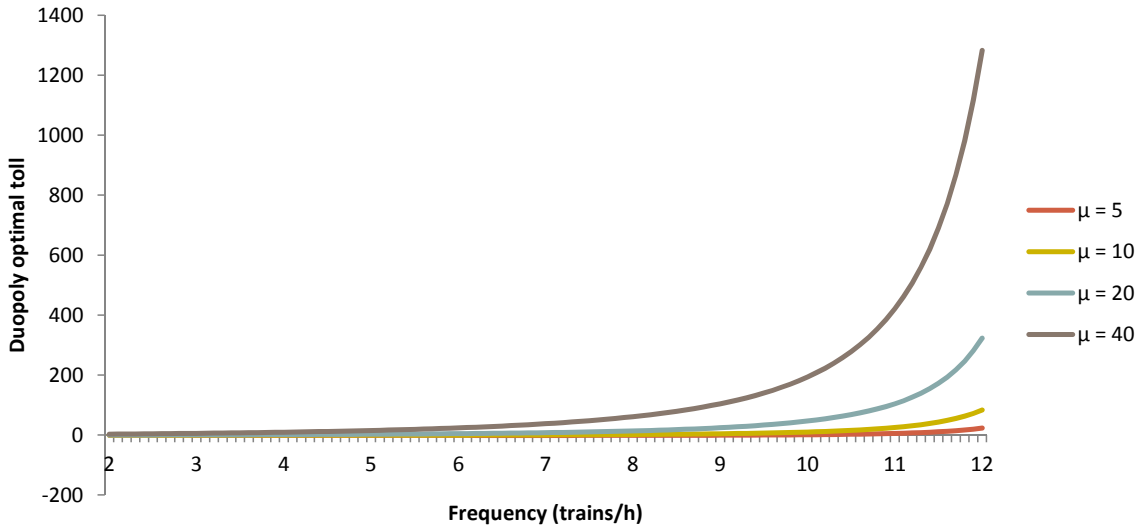

Figure 4.8: A change on the average initial delay: Impact of frequency on the duopoly optimal toll.

with a "Mohring effect"). In contrast, on rail lines with high market power, even in the presence of congestion externalities, the optimal toll should be negative to avoid worsening the initial distortion.

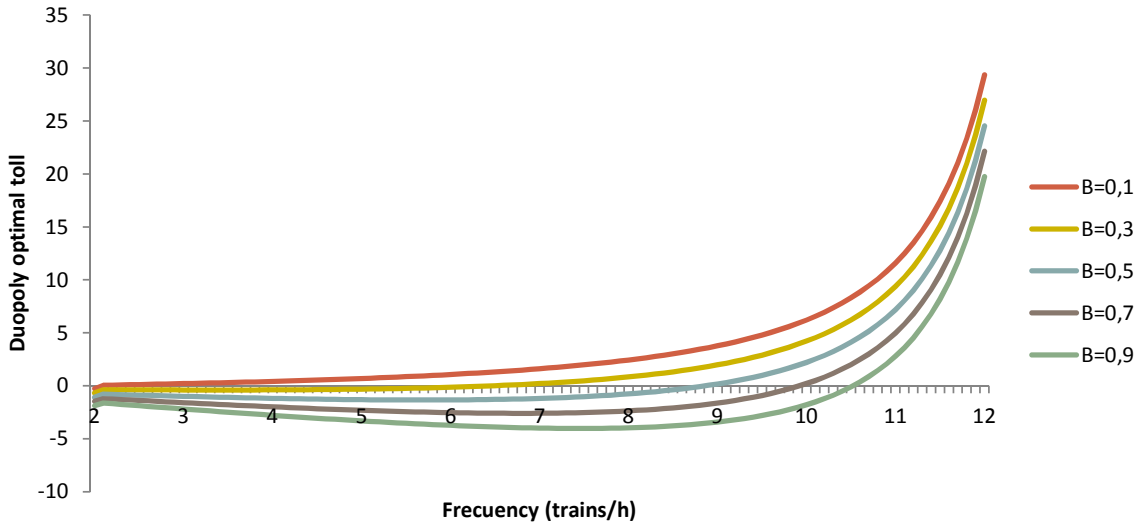

Figure 4.9: A change in the inverse demand sensitive parameter: Impact of frequency on the duopoly optimal toll.

4.5 Conclusion

Chapter 4 proposes a global vision of the railway capacity constraint problem, by considering an equilibrium supply-demand model under different market conditions.

The aim of this chapter is to verify and discuss whether the theoretical conclusions on other modes of transport applied to tackling rail capacity constraints in the short-term are also verified in the railway sector.

The results show that in some situations an optimal congestion toll can be considered as an optimal tool for dealing with congestion as in other transport modes. Nevertheless, the specificities of the railway sector suggest that the toll must be implemented under specific conditions in order to be optimal. As in the air sector, railway companies have market power and internalise part of the congestion externalities. Consequently, the estimation of optimal congestion toll must consider the market power effect.

Besides, as highlighted throughout this research, the particularity of the railway sector and this research is that an additional passenger can also generate a positive externality and not only a negative one. In some situations, when the expected schedule delay effect or Mohring effect is identified, the regulator should subsidise the service (implementing a negative toll) in order to encourage the TOC to provide an additional service.

To sum up and considering all these specificities, in the short-run an optimal congestion toll in the railway sector must take into account the firm's market power, the effect of particular externalities developed (positive or negative) and the initial quality of the service. Without these considerations, implementing a toll could create greater distortions than the non-regulated scenario.

Some extensions of the model can be considered for further research. Firstly, it will be interesting to analyse the conclusions of the supply-demand model with a different generalised user cost function, like the extension presented in the appendix B, or with dynamic congestion.

Other logical subsequent paths include considering other models of train operator company behaviour, different from the Cournot game. As in the literature on airlines, it is important to consider whether empirical research confirms Cournot competition in railways or whether another type of behaviour model like Bertrand or Stackleberg competition should be considered.

From the theoretical viewpoint, and by analogy with the air sector, it will be also interesting to analyse whether the policy recommendations for one kind of competition

are still applicable to a different one. For example, in the airline scenario, Silva and Verhoef (2013) investigated and compared whether the recommendations of considering market power internalisation in the congestion toll are the same under various types of competition. They concluded that in a differentiated Bertrand duopoly case, the amount of congestion that airlines internalise is smaller than the classical example of Cournot competition.

For future research, it would be a natural step to move to the long term perspective and the correlation between optimal pricing and capacity investment, via the self-financing theorem (Mohring and Harwitz, 1962). In the perspective of a long term railway analysis, it will be interesting to bear in mind the lessons drawn by airlines on this question.

The relationship between self-internalisation and congestion toll level also raises a new question on airport investment and capacity. Indeed, if we consider that firms are not atomistic and that they have some market power, the welfare-optimal congestion toll may be reduced, in order to adapt it to market power distortion. Thus the literature on airlines analysed whether the self-financing infrastructure theorem was still applicable in a non-atomistic scenario, and concluded that cost recovery cannot be achieved in a market power scenario (Brueckner, 2002; Zhang and Zhang, 2006). On the other hand, Verhoef and Mohring (2009), considered that the Mohring-Harwitz theorem can be applied, even in a non-atomistic world, if we consider that in the case where operators have market power, they also have an incentive to contribute privately to capacity investment.

To conclude, future research could focus on capacity and investment issues in the long term, by considering the impact of market power and different kinds of competition on traditional short and long term policy conclusions. These theoretical studies must be completed by empirical developments in order to determine the specificities of railway market competition compared to air transport.

Conclusions and policy recommendations

Conclusion and Policy Recommendations

The issue of capacity constraint and its optimal management is a major challenge for the French railway network. Up to now, capacity constraints in the railway sector have been viewed in a very compartmentalised manner. On the one hand, operational research was performed in view to optimising the timetable from the technical standpoint. On the other hand, and in the European regulatory context, we observed increasing economic consideration of a mainly theoretical nature that views railway congestion as a negative externality (similar to other modes of transport) and questioning of the pertinence of pricing as a corrective measure. However, there was no global approach of railway congestion linking optimal responses to adjustments between supply and demand that generate capacity constraints.

In the introduction to this dissertation we postulated that the railway, a system traditionally associated with increasing returns to scale, had entered a decreasing returns to density zone. New forms of mobility with a concentration of traffic at very specific times and places have underlined a deterioration of regularity and thus an increase in average costs for users above a certain threshold of traffic. In addition, from the standpoint of infrastructure, investments in hubs of congestion have above all been made in and around large cities, where construction costs are very high (cost of land, the need for underground tracks, construction sites difficult to access, etc.). Thus returns to scale will be decreasing for considerable increases in capacity.

Main Results of the Research

Throughout this research, we have attempted to precisely describe the technical and economic elements that characterise the issue of railway capacity constraints as a whole, and which provide us with a global view of the issue.

Technical and economic perspectives

Firstly, the definition of capacity constraints was examined from an engineering perspective via timetable design, a key element in matching supply with demand for programmed modes of transport. Better knowledge of timetable construction methods led to highlighting the implicit trade-offs between the capacity supplied and service quality in terms of reliability.

The analysis of the notion and measure of capacity constraint from the economic angle is a subject that has been dealt with extensively for other modes such as road and air transport. Economic studies relating to other modes of transport provided us with useful lessons for formalising the issue for railways (little studied up to now) and adapting their transposition to the railway industry.

A generalised user cost function

The combination of these two visions, the technical vision of the engineer and the economic vision developed for other modes of transport, allowed us to formulate a microeconomic model of the generalised cost of the user, taking into account the characteristics of timetable construction specific to railways. The modelling highlighted the dual effect of a higher frequency of rail traffic, with an impact on the expected schedule delay cost (Mohring effect) on the one hand, and the random delay cost linked to the intensive use of the network on the other.

The definition of the generalised cost function specific to railways that objectivises the trade-offs (tacit until now) between the capacity supplied and service quality in terms of reliability, leads to questioning the balance between supply and demand according to different market structures that define optimal prices and quantities.

Correcting externalities

This analysis therefore permits determining the conditions under which, and the objectives to be pursued by, the public authorities (as regulator) to adjust the shortcomings of

the decisions made by private operators. Consequently, the analysis developed showed that, under certain conditions, the regulator can be led to promote charging for capacity constraints to internalise the external effects generated and send the right price signals to economic operators.

As set out in the last chapter of this dissertation, using congestion pricing as a measure to correct the externalities linked to capacity constraints depends on several parameters:

- **The presence of a monopoly on the market**: A rational and perfectly discriminating monopoly determines the efficient allocation of capacities and fully internalises phenomena linked to capacity constraints (Mohring or delay effect). The justification of public intervention in the case of a monopolistic company is linked to the natural inefficiency of the monopoly (price mark-up higher than the marginal cost) and not to the absence of internalisation of capacity constraints.

- **Competition between railway operators**: Congestion pricing cannot therefore be justified except in a situation where there are several operators (at least a duopoly) on the market. In practice, the railway infrastructure supports different activities (regional transport, long distance transport, freight transport, etc.) and the combination of the different demands made by these activities in relation with each other may lead to them to reaching or exceeding the capacities of the system and thus cause congestion.

  Although most of these different services are provided by the same historic operator, the decision regarding the number of trips demanded by the activity is made in an uncoordinated way at different moments (by regional and national transport organisation authorities for activities subject to agreements, and by the operators for commercial activities). In the case of the railway infrastructure, the competition between activities can also be considered to be the source of externalities.

- **The market power of operators:** As in the air transport sector, congestion pricing justified by not taking account of externalities must be reduced as a function of the market power of companies. Indeed, a certain market power, for example linked to a low level of intermodal competition, gives an operator the opportunity to discriminate through its prices. If congestion pricing neglected this market power, its application would lead to an additional distortion of the market.

- **The initial level of quality in terms of the frequencies supplied and reliability:** An additional frequency on a line does not generate the same externality

as a function of the type of incidents at the outset (initial degree of delay) or the frequency initially supplied. For the same level of frequency, a line with a higher probability of incident would generate a higher externality than that of a line with a good level of reliability. In certain cases, an additional frequency generates a positive externality (Mohring effect), thereby justifying a subsidy to encourage increased production rather than additional pricing.

A wider vision of the optimal management of capacity constraints

In this approach, the correct signal-price resulting from this full analysis is based on a static vision of the question, and is only pertinent for an instantaneous, short term view. However, it should be emphasised that the public authorities (or the regulator) must not separate the question of congestion pricing from the other components of the capacity constraint problem. The pricing proposed in this research, which represents an optimal tool for solving inefficiencies, is based on the hypothesis according to which the network's dimensioning and reliability are optimal and fixed in the short term. The global analysis of capacity constraints should be seen in the long term, by taking into account the cost of increasing capacity, the variation of its parameters and its impact on price recommendations.

Apart from these long term considerations, which will be dealt with in more detail below, this work showed that the regulator (in the broad meaning) should not support congestion pricing without ensuring that the infrastructure manager allocates capacity as efficiently as possible. Congestion pricing as described in chapter 4 is subject to the optimal allocation of capacity, based on detailed knowledge of the generalised cost function of the user obtained by the infrastructure operators/manager, not forgetting their own costs. If these conditions are not taken into account (poor allocation of capacities through lack of knowledge of the user cost function), implementing congestion pricing may lead to a suboptimal situation and to a loss of value for the public authority. Let us imagine, for example, an infrastructure manager greatly averse to delays, which neglects the value of frequency for the user, and which overestimates margin times by limiting capacity. If the regulator decides to authorise congestion pricing, increasing the price would encourage a reduction of the frequencies demanded below the optimal frequency. Congestion pricing in a framework of overestimated margins would result in the under-utilisation of the available optimal capacity.

In brief, this research emphasises the importance of the components of the generalised

cost function of the user. Detailed knowledge of this function is required so that the infrastructure manager objectivises the trade-offs made in the timetable construction process, and justifies the pertinence and optimality of possible congestion pricing to the public authorities.

Nonetheless, in order for the short term recommendations to remain optimal in a long term economic perspective, the mismatch between supply and demand linked to congestion must be considered at the scale of the railway system, that is to say by comparing the final demand (passengers or goods) with the railway supply of available seats, frequencies, reliability and dimensioning.

Towards the Implementation of Pricing in the Present Regulatory Framework

As described in this dissertation, under certain conditions, implementing congestion pricing can be considered as a measure to correct externalities linked to capacity constraints. Economic intuition pleads for a global view of the railway system when analysing capacity constraints, also reflected in the European and national legal framework.

The European standpoint

European directive 2012/34 determines the legal framework of the link between these parameters explicitly.

Thus article 31 of the European directive authorises the infrastructure manager to apply *"a charge which reflects the scarcity of capacity of the identifiable section of the infrastructure during periods of congestion"* In this regulatory framework, pricing linked to capacity constraint can be applied if the infrastructure manager has formally declared the line or section of line of the infrastructure saturated beforehand.

Article 47 of this directive describes the regulatory provisions regarding infrastructure saturation. To take into account the transposition of directive 2012-34, article 26 of decree no. 2003-194 amended in August 2015, considers that the infrastructure manager must declare a section of the infrastructure saturated *"Where, after coordination of the requested train paths and consultation with applicants, it is not possible to satisfy requests for infrastructure capacity adequately(...)"*. The regulations add *"This shall also be done for infrastructure which can be expected to suffer from insufficient capacity in the near*

future" The particularity of this process transposed into national law from European law is that saturation can be "observed" or be "foreseeable".

According to article 50 of the European directive, where infrastructure has been declared to be congested,the infrastructure manager shall carry out a capacity analysis within a period of 6 months. The aim of this analysis is to determine the causes of the saturation and propose measures to correct them. According to the interpretation made throughout this research, this analysis of capacity should allow the infrastructure manager to justify at this point their trade-offs between frequency and regularity and prove, for example that if a train path is refused, that it is in the interest of the public authority to avoid degrading a certain level of quality service. From the economic standpoint, the justification of an unfavourable response to request for capacity should be based on the ability to demonstrate that a prior optimal allocation of available capacities exists.

Lastly, as stated in article 51 of the directive, within six months following the capacity analysis, the infrastructure manager must propose a capacity-enhancement plan which may be subject to prior approval by the Member State. This plan must set out *"the reasons for the congestion, the likely future development of traffic, the constraints on infrastructure development and the options and costs for capacity enhancement"*. The application of an additional charge during periods of saturation is subject to the presentation and implementation of actions determined in the capacity-enhancement plan.

From the regulatory standpoint, the infrastructure manager shall cease to levy any charge if it does not implement the actions of the capacity reinforcement plan. Thus the directive establishes a link between the additional income received by the infrastructure manager during periods of saturation and a policy to reinforce capacity (in which, among other things, investments in capacity can be found). In this way, it establishes a relation between short term instruments (pricing), and long term ones (investment and other measures taken to reinforce capacity) used to regulate railway capacity constraints.

In other European countries, it appears that half the railway infrastructure managements had declared some infrastructure to be congested. On the contrary, the perimeter of the sections impacted by this declaration differs from country to country and, up to now, the application of an additional charge associated with a capacity constraint has been little used as a lever of action. As mentioned, current consideration concerning the implementation of a congestion declaration procedure driven by European infrastructure managers, and its associated pricing, is now in a preliminary phase. In the French case, these considerations participate in a progressive approach by SNCF Réseau with the de-

sire of transparency and objectification of the activities linked to capacity constraints, from the economic standpoint.

The French case

Although the legal framework provides the possibility of declaring saturated infrastructures, at present the French infrastructure manager has never opted to use it. Up to now, setting up a congestion declaration procedure has been complicated from the practical angle and superfluous since the coordination process has enabled settling every conflict. However, certain lines and hubs (i.e. the high speed line between Paris and Lyon and Lyon railway hub (NFL), the Montpellier-Perpignan corridor and the Marseille-Nice line) are prone to saturation in the medium/long term if actions are not taken.

In the framework of French regulation, recommendation no. 2016-014 by Arafer (French rail and road regulatory body) relating to the Network Statement for the Service Schedule 2017[2], recommended that SNCF Réseau should use this congestion declaration procedure provided in directive 2012/34/UE and its transposition into national law when pertinent.

SNCF Réseau has since worked on the possibility of implementing the congestion declaration procedure set out in the directive and the pricing that could be linked to it for the operating year 2018. The proposal submitted to the regulator by SNCF Réseau is based on the capacity allocation process calendar[3] in force, and makes the clear distinction between a declaration of foreseeable saturation and observed saturation. It also proposes, eventually, a flat rate price system in the case of a declaration of predictable saturation, aimed at encouraging demanders of train paths to change their behaviour. This system (regarding its pricing aspect) will be proposed for testing during the first year (Service schedule 2018).

The rationale running through the European legal framework is one of opening out to competition. The regulator's requests for the clarification and justification of procedures will no doubt lead SNCF Réseau to investigate these issues further and consider all the related stakes and incentives in the years to come. Therefore it appears important to fuel reflection with a few additional initial elements of analysis. From the regulator's viewpoint these elements lead to assuming that evaluating the cost of capacity constraint

[2]In its previous recommendations no.2012-005 and no. 2013-002 relating to Network Statement 2013 and 2014 and no. 2014-001 and no. 2015-003 relating to Network Statement 2015 and 2016, the regulatory body had already recommended that SNCF Réseau should use this procedure.

[3]The French graphic timetable construction process is detailed in Appendix E.

could involve different domains of access to the network (infrastructure design, train path allocation, pricing).

Further Policy Recommendations

Firstly, reflections on pricing policies should be associated with long term issues, including the question of the correct dimensioning of the infrastructure. Indeed, validation of congestion pricing without a policy for future investments could, for example, incite the IM to under invest in physical capacity. This raises the question whether, in its own interest, the IM would benefit from investing in capacity if this meant an additional cost and a reduction of its congestion revenues. It appears essential for regulation to tackle the problem of capacity constraint by considering every component and time horizon, to ensure a virtuous system of incentives. For example, for railway companies, information on the existence of saturation periods should occur sufficiently early for the incentive to be effective, so they could adjust their demand for capacity to the presence of a congestion charge.

The short term optimum

Economic theory has given great attention to the link between optimal short term pricing and long term investments in capacity. As described in chapter 2 of this research, under certain conditions, i.e. optimal, revenue from congestion covers all the expenses associated with investments in capacity. Apart from the question of self-financing investments in capacity (subject to restrictive conditions poorly adapted to the reality of railways), it is interesting to examine in detail the link between these two variables under the prism of economics.

Optimisation of the trade-off between reliability and railway supply leads to a short-term optimum: the global cost of congestion is minimised and the user surplus is maximised. For all that, this short term optimum includes a share of residual congestion (users not travelling at the time desired and those delayed by the congested infrastructure). This congestion is optimal in the short term, but achieving a long-term optimum would require that the cost of this residual congestion remains lower than the cost of increasing the infrastructure's capacity, which would permit reducing it.

Increasing capacity

Indeed, congestion pricing generates short term revenues that will be used to finance (all or some of the investments in future capacity). Optimally, it would be preferable to invest in increasing capacity when the marginal cost of congestion (for a given capacity) is higher than the marginal cost of increasing capacity. This condition raises two questions:

Firstly, it assumes that the network manager (or regulator) is capable of evaluating the cost of the capacity in previous socioeconomic evaluations. However, we know that up to now, socioeconomic evaluations performed in the transport sector were mainly based on the value of the time saved. This methodological framework was very well adapted for justifying the construction of highways, then railway lines. But they come up against several difficulties when evaluating investments to create capacity, despite the fact that a change of methodology (Rapport Quinet, 2013) has emerged with the offer of new methodological tools for evaluating these development projects.

Secondly, this condition necessarily leads to temporary imbalances since the real world is characterised by indivisibilities. As indicated by Hau (1998), the optimal sequence of the decision-making process is first to set out a policy for implementing a price at marginal social cost and then plan future adjustments of capacity in relation to future demand and the pricing policies formulated. In the case of a system with indivisibilities like railways, adjustments between pricing and investment will not be made automatically. Initially, it is possible that, given the difference in time between the decision to make an investment and its effective realisation, congestion pricing will be introduced despite the existence of investments in capacity at costs lower than that of the residual capacity. In parallel, after the investment had been realised, the new dimensioning of capacity may eliminate the residual congestion that justified congestion pricing, and it should be stopped. The existence of this type of imbalance argues in favour of smoothing congestion pricing to send coherent signals on prices to the actors over a long period. In this case the charge should be reduced during the period when the infrastructure is under-dimensioned, provided that the charge can be extended once the investment has been made.

Nonetheless, smoothing congestion pricing can interfere with the objective of the right price signal in the short term if the operators cannot clearly identify the cost of the externalities generated by their private decisions. Finding a balance between the incentive effect of pricing for the optimal use of the infrastructure in the short term and its technical feasibility in a railway sector in which investments have a long lifespan is an open and complex question that requires in-depth consideration.

Other components linked to capacity constraints

Furthermore, other components such as performance and efforts linked to improvement that were considered fixed until now must be incorporated in a general long-term vision. As with investments, this vision is also reflected in regulatory texts. The recitals set out in directive 2012/34/UE state that *"It is desirable for railway undertakings and the infrastructure manager to be provided with incentives to minimise disruption and improve performance of the network"*. Thus article 35 of the directive specifies that a performance scheme *" may include penalties for actions which disrupt the operation of the network, compensation for undertakings which suffer from disruption and bonuses that reward better-than-planned performance"*.

In the framework of our analysis, capacity is considered as allocated efficiently in the short term given the initial disruptions observed. Nevertheless, if the infrastructure manager and the railway companies do not make the necessary efforts to minimise the number of disruptions in the long term, short term policies will fail to be efficient. Consequently, it is necessary to forge a link with the performance scheme designed to encourage the reduction of disruptions, whether they concern the infrastructure or the rolling stock. Improvements of performance can be translated in a good level of investment in the network's reliability or in a more efficient treatment of disruptions by the infrastructure manager and the railway companies.

Towards a Systemic Vision of Capacity Constraints

To conclude, railway congestion should not be simplified to a relation between the number of trains and delays or the introduction of a price. In the increasingly regulated world of rail transport, with more open and concerted decision-making procedures, the analysis of capacity constraints must stem from the analysis of the system, linking all the decisions involving short and long term capacities, such as:

- the main trade-offs of the capacity allocation process between frequency and regularity;
- the stakes of congestion pricing;
- the mechanisms of encourage railway undertakings and the infrastructure manager to minimise disruption and improve the performance of the railway network(performance scheme);

- the definition of a good level of investment in capacity;
- the enhancement of projects that create capacity and/or improve the robustness of the graph in the socio-economic analysis.

Appendix

Appendix A

Detailed Econometric Analysis

As stated in the introduction, an extensive econometric analysis was conducted for the French railway network, (Pérez Herrero et al., 2014). This appendix details the data set and the numerical results of this research.

A.1 The Data Set

In Pérez Herrero et al. (2014), we use data from an internal database of SNCF Réseau in order to estimate the parameters presented in the Introduction. This internal database records traffic information in the French network, and notably the delays at each measuring point. The data provided by this database allow us to know precisely the performance (reliability rate and delay) of each line at each level of traffic.

The data is recorded by an automatic system which detects the train circulation and registers the traffic details concerning the train. These automatic measuring points are associated to the measuring points which are utilised for the construction of the schedule. The system allows obtaining, for each train which crosses a measuring point, the data presented in the following table.

However, railway lines have different characteristics. They have diverse uses (passenger trains or freight trains), different traffic densities (lines with heavy traffic or lines with low traffic) and varied levels of performance. For that reason, we have subdivided the French network in several groups of lines with similar characteristics.

In this classification, the network is divided in 4 categories depending on uses (freight, regional, national) and speed levels (high speed lines or not):

- High speed lines: routes with a speed higher than 250 kph

Variables	Description
Internal circulation number	Specific and unique number associated at each train
Circulation number	Number associated to a specific stopping pattern
Date/Hour	Date et real hour when the train crosses the measuring point
Week day	-
Timetable type	Determines the kind of stop: Origin, Passage, Arrival or Departure (for a stop in a train station) or Terminus
Time deviation	It is the deviation between the real time and the scheduled time (delay)
Statistical category	Informs about the train activity (HSL, regional activity, national activity, freight, etc.) and if the train is loaded or empty

Table A.1: Summary variables

- Intercity lines: routes between population centers mainly used by freight and passenger long distance trains.

- Regional lines: routes between suburbs, towns and cities, without special speed requirements, and mainly used by regional and commuters trains.

- Only freight lines: freight specific routes with no mixed traffic, and generally low traffic density.

At the same time, these categories are subdivided in subcategories depending on the traffic density (trains per weekday per route): high, medium or low traffic density.

The traffic is highly concentrated around several nodes of the networks. For example, we can observe lines with 15 trains per hour during the peak-hours period in some regional railway lines near Paris. By contrast, some local lines can only have one train per hour during the peak-hours. The varied traffic lines density emphasises that congestion would not emerge with the same intensity in the entire network.

In the study, we focused our analysis on 42 lines of the French railway network, with 3 measuring points for each line. The lines belong to these different groups of lines presented above. The dataset includes 6.4 million trains (i.e. 6.4 million observations). These lines have been assembled in 9 subgroups using the strategic segmentation. The dataset used in this research contains all train circulations in these lines during 2011.

A.2 Results

For this analysis, the variable traffic has to be defined. For each observation (each train recorded), we have obtained a level of traffic which equals the number of train scheduled in the same line and direction during the previous hour. Then, an econometrical analysis is pursued to measure the additional delay (in minutes) in a railway route due to an increase of one traffic unit (the marginal delay). As mentioned above, an additional train is likely to be delayed and to impose an additional delay on the next trains. The consequences of an additional train (direct and indirect effect) have been considered separately in our analysis, in order to assess the effect that an additional train generates on other trains. The indirect effect is the pure externality from an economist point of view whereas the indirect effect is internalised by the additional train.

Some of the parameters are directly computed using the data set. Some others are estimated with the econometric analysis, as described above. Two econometrical regressions are conducted in order the estimate the marginal cost of congestion (indirect effect) in minutes: the probit model which estimates the marginal effect of traffic on the probability of being late, and the linear model which estimates the marginal effect of an additional train on the expected delay.

The results of the econometric analysis are presented in table A.2. The regressions have been estimated separately for the 9 groups of lines. Table refRegressions results presents the results of the two regressions which allow to calculate the indirect effect defined in the Introduction.

$$\frac{\partial E(R_i)}{\partial Q_i} = \frac{\partial p(R_i^* > 0)?E(R_i \mid R_i^* > 0)}{\partial Q_i} = \frac{\partial p(R_i^* > 0)}{\partial Q_i} E(R_i \mid R_i^* > 0) + (R_i^* > 0) \frac{\partial E(R_i \mid R_i^* > 0)}{\partial Q_i} \tag{A.1}$$

The first column represents the average marginal effect of an additional train on the probability of being late, calculated using a probit model. It correspond to the parameter $\frac{\partial p(R_i^*>0)}{\partial Q_i}$. The second column represents the marginal effect of an additional train on the expected delay calculated using a linear regression. It corresponds to the parameter $\frac{\partial E(R_i|R_i^*>0)}{\partial Q_i}$.

Standard error are in parentheses($*p < 0.10, **p < 0.05, ***p < 0.001$)

These results can be interpreted as follows: for high speed lines, an additional train increases the probability of being late by 0.96 points and increases the expected delay by

Strategic Classification	Type of line	Probit	Linear regression
G1	High Speed	0.0096***(0.0024)	0.020**(0.017)
G2	Intercity lines	0.020***(0.00042)	0.49** (0.12)
G3	Intercity/Regional lines	0.013*** (0.00005)	0.10** (0.018)
G4	Intercity lines high traffic density	0.022*** (0.00024)	0.67** (0.073)
G5	Intercity lines low traffic density	0.018*** (0.00057)	0.67 (0.30)
G6	Intercitylines medium traffic density	0.010***(0.0011)	0.19 (0.14)
G7	Regional lines high traffic density	0.025*** (0.00024)	0.14** (0.024)
G8	Regional lines low traffic density	0.056*** (0.0064)	0.67** (0.31)
G9	Regional lines medium traffic density	-0.025*** (0.0024)	1.05 (1.10)

Table A.2: Regressions results.

0.20 minutes for the following trains. Moreover, these results show that for certain types of lines, the congestion is not statistically significant. It is the case of intercity lines. It not surprising since this group corresponds to low traffic group of lines.

Once these two regressions have been estimated, it is possible to compute the average direct effect, and the marginal effect by group. The results of these computations are presented in table A.3. This table can be interpreted as follows: for G4, an extra train generates 0.68 minutes of delay on the forthcoming trains.

Strategic Classification	Marginal Effect
G1	0.19
G2	0.47
G3	0.13
G4	0.68
G5	0.59
G6	0.18
G7	0.19
G8	0.67
G9	0.30

Table A.3: Congestion marginal cost.

In order to check the robustness of these results, some tests have been realised. A first

test is realised in order to verify the existence of the relationship with another definition of delay. The previous results considered a train delayed if delay was superior to zero. Nevertheless, the data shows that many trains have in fact little delays (less than 5 minutes). A little delay associated to a train could be a measure error in some points, so we decided to test our results using a different delay definition. Two tests have been done considering only delays superior to three and five minutes respectively. In both cases, even if the absolute value of the direct effect is different, the estimated relationships are significant and the hierarchy between lines does not change.

Strategic Classification	**Marginal Effect** (> 3)	**Marginal Effect** (> 5)
G1	0.22	0.20
G2	0.37	0.27
G3	0.12	0.10
G4	0.62	0.53
G5	0.52	0.46
G6	0.076	0.051
G7	0.14	0.082
G8	0.51	0.56
G9	-0.018	0.09

Table A.4: Robustness test

Until now we have considered that marginal effects are homogeneous between measuring points or lines in the same group. Some regressions analyses have been also conducted for several specific points The test shows that there exist some differences between measuring points and lines. In some measuring points the congestion effects are higher than in others sections of the network, but the effect remains significant from a statistical point of view.

These results therefore provide strong evidence of our intuitive idea: an additional train increases the probability of late trains. It means that there is a form of unexpected congestion in the railways. The direct effect is internalised by the supplementary train, but the indirect effect generates an external cost on other users.

Acknowledgements

We would like to thank SNCF Réseau which funded and supervised this study and G. de Muizon, E. Frot and A. Charpin from Microeconomix for their contribution.

Appendix B

An Extension on the Consumer Generalised Cost Function

In this PhD dissertation, we have not considered that travellers may account for potential delays in their scheduling. Travel time variability (VTTV) disrupts one's activity planning, like early or late arrivals at destination. A delayed arrival could reduce the schedule delay cost for (at least some of) those who would otherwise have arrived too early. In the extreme case, the delay may cause the cancellation of the final destination activity.

An extended analysis considers that users are perfectly informed and that they integrate random travel times in their departure time choice, as developed in this appendix. This is the result of a collaboration with Nicolas Coulombel (Assistant Professor, Paris-Est - LVMT).

As exposed in chapter 2, scheduling models assume that the utility of individuals depends on the time spent on each activity: home, travel and work. Noting t_D the train departure time, T the planned travel time, and $t_A = t_D + T$ the arrival travel time. Utility V_0 has the following form, introduced by Vickrey (1973) and later worked out by Tseng and Verhoef (2008):

$$V_0(t_D, T) = \int_{t_H}^{t_D} h(t)dt + \int_{t_D+T}^{t_w} w(t)dt \tag{B.1}$$

where h is the marginal utility of time at home, and w the marginal utility of time at work. The constant t_H denotes the start of the day and t_w the work end time. Furthermore, we assume without loss of generality that h and w intersect at $t = 0$, so that the preferred arrival time t^* (in case of instantaneous travel, i.e for $T = 0$) is normalised to

0 in equation B.1. However, as exposed in chapter 3, users have heterogeneous preferred arrival time t^*. We therefore extend equation B.1 to the case of heterogeneous preferred arrival time as follows:

$$V(t_D, T, t^*) = V_0(t_D - t^*, T) = \int_{t_H}^{t_D - t^*} h(t)dt + \int_{t_D - t^* + T}^{t_w} w(t)dt \tag{B.2}$$

As stated in this research, in some cases, stochastic delays can increase the schedule travel time and delay propagation will depend on traffic density, which depends on the headways H between two trains. Stochastic delays increase travel time and can be written as: $\tilde{\mathrm{d}} = \mu(H) + \sigma(H)\tilde{\mathrm{x}}$, where $\mu(H)$ is the mean random delay and $\sigma(H)\tilde{\mathrm{x}}$ the standard deviation of the random delay. Given that travel time $\tilde{\mathrm{T}} = T_0 + \tilde{\mathrm{d}}$ is now stochastic, utility is also stochastic and can be rewritten as:

$$\tilde{\mathrm{V}}(t_D, T_0, t^*, H) = \int_{t_H}^{t_D - t^*} h(t)dt + \int_{t_D - t^* + T_0 + \mu(H) + \sigma(H)\tilde{\mathrm{x}}}^{t_w} w(t)dt \tag{B.3}$$

Faced with uncertain travel conditions, users choose their train departure time t_D based on the expected utility:

$$V(t_D, T_0, t^*, H) = E\left[\tilde{\mathrm{V}}(t_D, T_0, t^*, H)\right] \tag{B.4}$$

Let us note $\delta_{i-1,i}$ the critical value for which a user with preferred arrival time $t^* = \delta_{i-1,i}$ is indifferent between train $i-1$ and train i. As trains are separated by a constant headway H, and users have uniform preferred arrival times, we have $\delta_{i,i+1} = \delta_{i-1,i} + H$. Furthermore, the average utility for users is independent of the train considered, and is given by:

$$U(H) = \frac{1}{H} \int_{\delta_{i-1,i}}^{\delta_{i-1,i}+H} E\left[\tilde{\mathrm{V}}(t_D, T_0, t^*, H)\right] dt^* \tag{B.5}$$

We now study the influence of a marginal change in headway on the average expected utility of train users:

$$U^{'}(H) = \frac{1}{H}\frac{\partial(\delta_{i-1,i}+H)}{\partial H}\left[V(t_D,T_0,\delta_{i,i+1},H) - V(t_D,T_0,\delta_{i-1,i},H)\right] + \frac{1}{H}V(t_D,T_0,\delta_{i,i+1},H) - \frac{U(H)}{H} + \frac{1}{H}\int_{\delta_{i-1,i}}^{\delta_{i-1,i}+H} E\Big[w(t_D+T_0+\mu(H)+\sigma(H)\tilde{\mathrm{x}}-t^*)\mu^{'}(H) - \tilde{\mathrm{x}}w(t_D+T_0+\mu(H)+\sigma(H)\tilde{\mathrm{x}}-t^*)\sigma^{'}(H)dt^* \quad \text{(B.6)}$$

which can be rewritten as:

$$U^{'}(H) = \frac{V(t_D,T_0,\delta i,i+1,H) - U(H)}{H} - \frac{\mu^{'}(H)}{H}\int_{\delta_{i-1,i}}^{\delta_{i-1,i}+H} E\left[w(t_D+\tilde{\mathrm{T}}-t^*)\right]dt^* - \frac{\sigma^{'}(H)}{H}\int_{\delta_{i-1,i}}^{\delta_{i-1,i}+H} E\left[\tilde{\mathrm{x}}w(t_D+\tilde{\mathrm{T}}-t^*)\right]dt^* \quad \text{(B.7)}$$

The first right hand term of equation B.7 corresponds to the increase in expected schedule delay related to H (the difference between the marginal user utility and the average user utility). The second and third right hand terms correspond respectively to the travel time benefit and to the reliability benefit of the marginal change in service headway.

Based on the findings of the literature on the value of travel time savings ($VTTS$) and the value of travel time variability ($VTTV$)[1] the integral (divided by H) in the second right hand term can be interpreted as the average $VTTS$ of users of train i, and the third term as the average $VTTV$ of users of train i.

This new formulation of the user utility reflects more precisely the travellers' behaviour when choosing their train service departure time. It allows us to consider the possibility that users can anticipate random travel times in their departure train time decision. Nevertheless, it seems quite complicated to solve analytically equation B.7. One of the objectives of this research is to find an optimal frequency from a consumer perspective, but not necessarily an optimal departure time for users (although anticipating delay is seemingly a major issue). In that perspective, we opted for considering travellers as naive or occasional, possibly experiencing an expected schedule cost and a random delay cost, but not anticipating any interaction between both costs.

[1] See Jenelius (2012) for instance

Appendix C

Numerical Value of the Parameters

This appendix recapitulate the numerical value of the parameters employed in this dissertation. As stated, it is not the purpose of this work to describe the real-life rail industry. The parameters therefore need not to correspond to real life values.

Table C.1 summarizes the multipliers values used in the simulations for the comparative statics and sensibility analysis in chapter 3.

β	γ	δ	μ	σ
Schedule delay early multiplier	Schedule delay late multiplier	Lateness penalty multiplier	Average initial delay	Delay's standard deviation
2.5	3	6	5 min	3 min

Table C.1: References values for the multipliers values

Table C.2 describes the values for the demand characteristics and TOC's costs used in chapter 4.

A	B	c_O^f	c_O^d	F_O	
Maximum reservation price	Demand sensitive parameter	Operating cost per train	Congestion cost for TOCs	Fixed operating cost	Lagrange multiplier
12	-0.7	2	1.5	3	0.3

Table C.2: Parameters values for demand and TOC's cost function

Appendix D

Notations

Variable	Definition *(unit)*
f	Number of trains/unit time
N	Total demand
T	Unit time
z	Number of users per unit time
GC_0	Monetized time cost function for a user
C_T	Planned travel time
C_E	Expected schedule delay cost
C_R	Random delay cost
t^*	Preferred arrival time
H	Headway
α	Travel time value
β	Schedule delay cost of arriving early
γ	Schedule delay cost of arriving late
t^i	Passenger indifference arrival time
k	Number of users per train
H_{min}	Minimum technical headway
H_{extra}	Buffer time
f_{max}	Maximal capacity of the line
$d_{1,i}$	Initial or primary delay
μ	Initial average delay
σ	Standard deviation of delay
δ	Random delay time cost

Table D.1: Summary list of main notations in chapter 3

Variable	Definition *(unit)*
$D(N)$	Inverse aggregate demand
A	Maximum reservation price
B	Demand sensitive parameter
GC	Generalised user cost
p	Fare
c_O^f	Operating cost per train
c_O^d	Congestion cost for TOCs
F_O	Fixed operating cost
τ_O^f	Toll per train
λ	Lagrange multiplier
M	Number of firms
m	Market shares of firms

Table D.2: Summary list of main notations in chapter 4

Appendix E

The French Graphic Timetable Construction Process

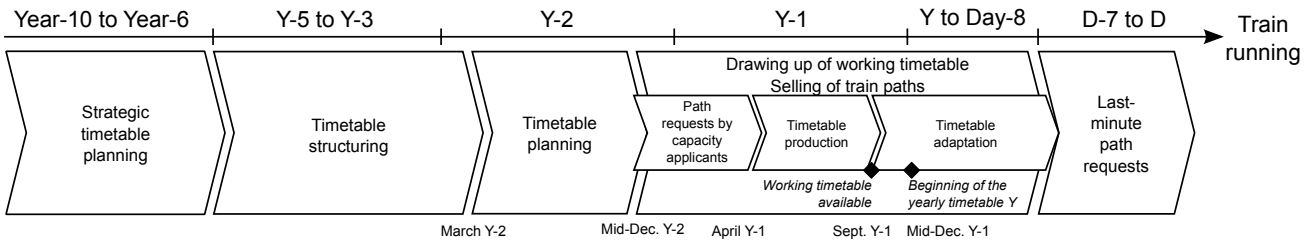

Figure E.1: The French graphic timetable construction process. Source: Morvant (2015)

Bibliography

Abril, M., Barber, F., Ingolotti, L., Salido, M. A., Tormos, P., and Lova, A. (2008). An assessment of railway capacity. *Transportation Research Part E: Logistics and Transportation Review*, 44(5):774–806.

Arnott, R., De Palma, A., and Lindsey, R. (1988). *Schedule delay and departure time decisions with heterogeneous commuters*. Number 1197.

Arnott, R., De Palma, A., and Lindsey, R. (1990). Economics of a bottleneck. *Journal of urban Economics*, 27(1):111–130.

Arnott, R., De Palma, A., and Lindsey, R. (1993). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review*, pages 161–179.

Arnott, R. and Kraus, M. (1998). Self-financing of congestible facilities in a growing economy. *1998) Topics in Public Economics: Theoretical and Applied Analysis Cambridge University Press, Cambridge UK*, pages 161–184.

Arnott, R. and Kraus, M. (2008). Congestion. In *The New Palgrave Dictionary of Economics*, pages 110–113. Nature Publishing Group, Basingstoke, 2 edition.

ARUP & Network Rail (2013). Recalibrating the Capacity Charge for CP5. Technical report.

Barber, F., Abril, M., Salido, M. A., Ingolotti, L., Tormos, P., and Lova, A. (2007). *Survey of automated systems for railway management*. F. Barber, M. Abril, MA Salido, L. Ingolotti, P. Tormos, A Lova.//Department of Computer Systems and Computation, Tech-nical University of Valencia, DSIC-II/01/07.–2007.

Basso, L. J., Jara-Díaz, S. R., and Waters, W. (2011). Cost functions for transport firms. *ResearchGate*, pages 273–297.

Bates, J., Polak, J., Jones, P., and Cook, A. (2001). The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review*, 37(2–3):191–229.

Batley, R., Dargay, J., and Wardman, M. (2011). The impact of lateness and reliability on passenger rail demand. *Transportation Research Part E: Logistics and Transportation Review*, 47(1):61–72.

Batley, R. and Ibáñez, J. N. (2012). Randomness in preference orderings, outcomes and attribute tastes: An application to journey time risk. *Journal of choice modelling*, 5(3):157–175.

Baumol, W. J., Panzar, J. C., and Willig, R. D. (1982). Contestable markets and the theory of industry structure.

Becker, G. S. (1965). A Theory of the Allocation of Time. *The economic journal*, pages 493–517.

Benezech, V. and Coulombel, N. (2013). The value of service reliability. *Transportation Research Part B: Methodological*, 58:1–15.

Boiteux, M. (1956). Sur la gestion des monopoles publics astreints à l'équilibre budgétaire. *Econometrica, Journal of the Econometric Society*, pages 22–40.

Braeutigam, R. (1999). *Learning about Transport Costs.* Brookings Institution Press.

Brueckner, J. K. (2002). Airport congestion when carriers have market power. *American Economic Review*, pages 1357–1375.

Brueckner, J. K. (2005). Internalization of airport congestion: A network analysis. *International Journal of Industrial Organization*, 23(7):599–614.

Burdett, R. L. and Kozan, E. (2006). Techniques for absolute capacity determination in railways. *Transportation Research Part B: Methodological*, 40(8):616–632.

Cantos, P. (2001). Vertical relationships for the European railway industry. *Transport Policy*, 8(2):77–83.

Cantos, P. and Maudos, J. (2001). Regulation and Efficiency : the case of European railways. *Transportation Research Part A: Policy and Practice*, 35(5).

Carey, M. (1999). Ex ante heuristic measures of schedule reliability. *Transportation Research Part B: Methodological*, 33(7):473–494.

Carlin, A. and Park, R. E. (1970). Marginal Cost Pricing of Airport Runway Capacity. *The American Economic Review*, 60(3):310–319.

Carrion, C. and Levinson, D. (2012). Value of travel time reliability: A review of current evidence. *Transportation research part A: policy and practice*, 46(4):720–741.

Caves, D. W., Christensen, L. R., and Swanson, J. A. (1981). Productivity Growth, Scale Economies, and Capacity Utilization in U.S. Railroads, 1955-74. *The American Economic Review*, 71(5):994–1002.

Caves, D. W., Christensen, L. R., and Tretheway, M. W. (1984). Economies of density versus economies of scale: why trunk and local service airline costs differ. *The RAND Journal of Economics*, pages 471–489.

Chen, B. and Harker, P. T. (1990). Two moments estimation of the delay on single-track rail lines with scheduled traffic. *Transportation Science*, 24(4):261–275.

Chu, X. (1995). Endogenous trip scheduling: the Henderson approach reformulated and compared with the Vickrey approach. *Journal of Urban Economics*, 37(3):324–343.

Cohen, Y. (1987). Commuter welfare under peak-period congestion tolls: who gains and who loses? *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti*, pages 239–266.

Commission, E. (2014). Fourth report on monitoring development of the rail marlet (RMMS). Technical report.

Daniel, J. I. (1995). Congestion pricing and capacity of large hub airports: A bottleneck

model with stochastic queues. *Econometrica: Journal of the Econometric Society*, pages 327–370.

DCF-DPS Supervision et Support. Réseau Ferré de France (2006). Détermination et confection des horaires. Technical report.

De Palma, A. and Fontan, C. (2001). Eléments d'analyse de la composante horaire des déplacements: le cas de la région Ile-de-France. *Les Cahiers Scientifiques du Transport*, 39:55–86.

De Palma, A. and Fosgerau, M. (2010). Dynamic and Static congestion models: A review.

De Palma, A. and Lindsey, R. (2001). Optimal timetables for public transportation. *Transportation Research Part B: Methodological*, 35(8):789–813.

De Palma, A. and Lindsey, R. (2007). Chapter 2 Transport user charges and cost recovery. *Research in Transportation Economics*, 19:29–57.

De Rus, G. and Román, C. (2006). Análisis económico de la línea de alta velocidad Madrid-Barcelona. *Revista de economía aplicada*, 42:35–79.

Dessouky, M. M. and Leachman, R. C. (1995). A simulation modeling methodology for analyzing large complex rail networks. *Simulation*, 65(2):131–142.

DGDI Bureau des Horaires, SNCF (2007). *Détermination et confection des horaires en zone dense de l'Ile de France*.

Dupuit, J. (1854). *Péages, Dictionnaire de l'Economie Politique, vol II*. Guillaumin–Coquelin.

Duron, P. (2013). Mobilité 21 - « Pour un schéma national de mobilité durable » -. Technical report, Ministère de l'écologie, du développement durable et de l'énergie.

Eurocontrol (2016). CODA Digest 2015 | Eurocontrol. Technical report.

Fosgerau, M. (2009). The marginal social cost of headway for a scheduled service. *Transportation Research Part B: Methodological*, 43(8):813–820.

Fosgerau, M. and Engelson, L. (2011). The value of travel time variance. *Transportation Research Part B: Methodological*, 45(1):1–8.

Fosgerau, M. and Karlström, A. (2010). The value of reliability. *Transportation Research Part B: Methodological*, 44(1):38–49.

Frank, O. (1966). Two-way traffic on a single line of railway. *Operations Research*, 14(5):801–811.

Gaver Jr, D. P. (1968). Headstart strategies for combating congestion. *Transportation Science*, 2(2):172–181.

Gibson, S., Cooper, G., and Ball, B. (2002). The evolution of capacity charges on the UK rail network. *Journal of Transport Economics and Policy*, pages 341–354.

Gleave, S. D. (2011). Impact assessment of revisions to Regulation 95/93. Technical report, European Commission (DG Move).

Greenshields, B. D., Channing, W., Miller, H., and others (1935). A study of traffic capacity. In *Highway research board proceedings*, volume 1935. National Research Council (USA), Highway Research Board.

Guiraud, L., Gayda, S., Cabrita, I., and Kroes, E. (2014). Value of travel time reliability on French high-speed and regional services. In *Transport Research Arena (TRA) 5th Conference: Transport Solutions from Research to Deployment*.

Haith, J., Johnson, D., and Nash, C. (2014). The case for space: the measurement of capacity utilisation, its relationship with reactionary delay and the calculation of the capacity charge for the British rail network. *Transportation Planning and Technology*, 37(1):20–37.

Haith, J. A. (2015). *Understanding the Relationship Between Capacity Utilisation and Performance and the Implications for the Pricing of Congested Rail Networks.* PhD thesis, University of Leeds.

Hansen, I. and Pachl, J. (2014). Railway timetabling and operations. *Eurailpress, Hamburg*.

Harker, P. T. and Hong, S. (1994). Pricing of track time in railroad operations: An internal market approach. *Transportation Research Part B: Methodological*, 28(3):197–212.

Hau, T. D. (1998). 3. Congestion pricing and road investment. *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*, page 39.

Higgins, A. and Kozan, E. (1998). Modeling train delays in urban networks. *Transportation Science*, 32(4).

Hotelling, H. (1929). Stability in Competition. *The Economic Journal*, 39(153):41–57.

Inrix (2015). Urban Mobility Scorecard Annual Report. Technical report.

Jansson, K. (1993). Optimal public transport price and service frequency. *Journal of Transport Economics and Policy*, pages 33–50.

Jenelius, E. (2012). The value of travel time variability with trip chains, flexible scheduling and correlated travel times. *Transportation Research Part B: Methodological*, 46(6):762–780.

Knight, F. H. (1924). Some fallacies in the interpretation of social cost. *The Quarterly Journal of Economics*, pages 582–606.

Kolm, S.-C. (1968). *La théorie économique générale de l'encombrement*. SÉDÉ IS.

Kontaxi, E. and Ricci, S. (2010). Techniques and methodologies for carrying capacity evaluation: comparative analysis and integration perspectives. *Journal of Ingegneria Ferroviaria*, 11:1051–1080.

Krueger, H. (1999). Parametric modeling in rail capacity planning. In *Simulation Conference Proceedings, 1999 Winter*, volume 2, pages 1194–1200. IEEE.

Landex, A. (2008). *Methods to estimate railway capacity and passenger delays*. Technical University of Denmark (DTU).

Li, M. Z. (2008). A generic characterization of equilibrium speed-flow curves. *Transportation Science*, 42(2):220–235.

Li, Z., Hensher, D. A., and Rose, J. M. (2010). Willingness to pay for travel time reliability in passenger transport: A review and some new empirical evidence. *Transportation Research Part E: Logistics and Transportation Review*, 46(3):384–403.

Lin, M. H. and Zhang, Y. (2016). Hub-Carrier Scheduling and Hub-Airport Congestion Pricing. *Available at SSRN 2715407*.

Lindsey, C. R. and Verhoef, E. T. (2000). Traffic Congestion and Congestion Pricing. Technical Report 00-101/3, Tinbergen Institute Discussion Paper.

Lévy-Lambert, H. (1968). Tarification des Services à Qualité Variable–Application aux Péages de Circulation. *Econometrica*, 36(3/4):564–574.

Mayer, C. and Sinai, T. (2003). Network effects, congestion externalities, and air traffic delays: Or why not all delays are evil. *The American Economic Review*, 93(4):1194–1215.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of public economics*, 3(4):303–328.

Mill, J. S. (1848). Principles of Political Economy With Some of Their Applications to Social Philosophy. 1857. *George Routledge and Sons, Manchester*.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *The American Economic Review*, pages 591–604.

Mohring, H. and Harwitz, M. (1962). Highway benefits: An analytical framework.

Morrison, S. A. (1986). Special Issue Road PricingA survey of road pricing. *Transportation Research Part A: General*, 20(2):87–97.

Morrison, S. A. and Winston, C. (2007). Another look at airport congestion pricing. *The American Economic Review*, pages 1970–1977.

Morrison, S. A., Winston, C., Bailey, E. E., and Kahn, A. E. (1989). Enhancing the performance of the deregulated air transportation system. *Brookings Papers on Economic Activity. Microeconomics*, pages 61–123.

Morvant, C. (2015). Challenges raised by freight for the operations planning of a shared-use rail network. A French perspective. *Transportation Research Part A: Policy and Practice*, 73:70–79.

Mosca, M. (2008). On the origins of the concept of natural monopoly: Economies of scale and competition. *The European Journal of the History of Economic Thought*, 15(2):317–353.

Motta, M. (2004). *Competition Policy: Theory and Practice*. Cambridge University Press.

Murali, P., Dessouky, M., Ordóñez, F., and Palmer, K. (2010). A delay estimation technique for single and double-track railroads. *Transportation Research Part E: Logistics and Transportation Review*, 46(4):483–495.

Nash, C. and Samson, T. (1999). *Calculating transport congestion and scarcity costs. Final report of the expert advisors to the high level group on infrastructure charging (Working Group 2)*. European Commission, Brussels, Belgium.

Nash, C. A. and Matthews, B. (2003). Rail infrastructure charges–the issue of scarcity. In *First Conference on Railroad Industry Structure, Competition and Investment, Toulouse, on*, pages 7–8.

Newell, G. F. (1987). The morning commute for nonidentical travelers. *Transportation Science*, 21(2):74–88.

Noland, R. B. and Small, K. A. (1995). *Travel-time uncertainty, departure time choice, and the cost of the morning commute*. Institute of Transportation Studies, University of California, Irvine.

Nuzzolo, A. and Russo, F. (1998). Departure time and path choice models for intercity transit assignment. *Travel Behaviour Research: updating the state of play*.

Oum, T. H., Waters, W. G., and Yu, C. (1999). A survey of productivity and efficiency measurement in rail transport. *Journal of Transport economics and Policy*, pages 9–42.

Pels, E. and Verhoef, E. T. (2004). The economics of airport congestion pricing. *Journal of Urban Economics*, 55(2):257–277.

Perennes, P. (2014). *Rail sector specificities and its liberalization process: the price signal issue*. Theses, Université Paris 1 PAnthéon Sorbonne.

Petersen, E. R. (1974). Over-the-road transit time for a single track railway. *Transportation Science*, 8(1):65–74.

Pigou, A. C. (1920). The economics of welfare. *McMillan&Co., London*.

Pérez Herrero, M., Brunel, J., and Marlot, G. (2014). Rail externalities: assessing the social cost of rail congestion. In *Transport Research Arena (TRA) 5th Conference: Transport Solutions from Research to Deployment*.

Quinet, E. (1997). Full social cost of transportation in Europe. *The Full Costs and Benefits of Transportation*, pages 69–111.

Quinet, E. (2003). Short term adjustments in rail activity: the limited role of infrastructure charges. *Transport Policy*, 10(1):73–79.

Rebeyrotte, E. (2016). Observatoire de la saturation ferroviaire entre Nîmes et Perpignan. Technical report, CGEDD.

Salop, S. C. (1979). Monopolistic Competition with Outside Goods. *The Bell Journal of Economics*, 10(1):141.

Samuelson, P. A. (1948). *Economics : An Introductory Analysis*. New York, Toronto, London: McGraw-Hill Book Coy.

Santos, G. (2004). *Road pricing: theory and evidence*, volume 9. Elsevier.

Santos, G., Verhoef, E., and others (2011). Road congestion pricing. In *A Handbook of Transport Economics*.

Silva, H. E. and Verhoef, E. T. (2013). Optimal pricing of flights and passengers at congested airports and the efficiency of atomistic charges. *Journal of Public Economics*, 106:1–13.

Small, K. A. (1982). The scheduling of consumer activities: work trips. *The American Economic Review*, pages 467–479.

Small, K. A. and Verhoef, E. T. (2007). *The economics of urban transportation*. Routledge.

SNCF (2001). Robustesse des graphiques de circulation. Technical report.

SNCF Réseau (2016). *Rail network statement. Appendix 8.1: Technical reference documents for train path construction*.

Transportation Research Board (2013). Transit Capacity and Quality of service Manual, 3erd edition. Chapter 8- Rail Transit Capacity. Technical report.

Tseng, Y.-Y., Rietveld, P., and Verhoef, E. T. (2012). Unreliable trains and induced rescheduling: implications for cost-benefit analysis. *Transportation*, 39:387–407.

Tseng, Y.-Y. and Verhoef, E. T. (2008). Value of time by time of day: A stated-preference study. *Transportation Research Part B: Methodological*, 42(7):607–618.

Union Internationale de Chemins de Fer (2004). UIC Code 406–Capacity. Technical report.

Verchere, P.-M. and Djellab, H. (2013). Robustesse et résilience: des plans de transport ferroviaires. *Revue générale des chemins de fer*, (233):6–26.

Verhoef, E. T. (2001). An integrated dynamic model of road traffic congestion based on simple car-following theory: exploring hypercongestion. *Journal of Urban Economics*, 49(3):505–542.

Verhoef, E. T. and Mohring, H. (2009). Self-financing roads. *International Journal of Sustainable Transportation*, 3(5-6):293–311.

Vickrey, W. (1973). *Pricing, metering, and efficiently using urban transportation facilities*. Number 476.

Vickrey, W. S. (1969). Congestion theory and transport investment. *The American Economic Review*, pages 251–260.

Villemeur, D., Billette, E., Ivaldi, M., Quinet, E., and Urdanoz, M. (2015). The Social Cost of Air Traffic Delays.

Walters, A. A. (1961). The theory and measurement of private and social cost of highway congestion. *Econometrica: Journal of the Econometric Society*, pages 676–699.

Wardrop, J. G. (1952). Some theoretical aspects of road traffic research.

Yuan, J. and Hansen, I. A. (2007). Optimizing capacity utilization of stations by estimating knock-on train delays. *Transportation Research Part B: Methodological*, 41(2):202–217.

Zhang, A. and Czerny, A. I. (2012). Airports and airlines economics and policy: an interpretive review of recent research. *Economics of Transportation*, 1(1):15–34.

Zhang, A. and Zhang, Y. (2006). Airport capacity and congestion when carriers have market power. *Journal of Urban Economics*, 60(2):229–247.

List of Figures

List of Tables